PRO GRADU -TUTKIELMA

**Ansa Lilja**

# Trajektorianalyysin soveltaminen tuotantoprosessin muodostaman aineiston mallinnukseen

TAMPEREEN YLIOPISTO
Informaatiotieteiden yksikkö
Tilastotiede
Kesäkuu 2015

Tampereen yliopisto
Informaatiotieteiden yksikkö
LILJA, ANSA: Trajektorianalyysin soveltaminen tuotantoprosessin muodostaman aineiston mallinnukseen
Pro gradu -tutkielma, 41 s., 6 liites.
Tilastotiede
Kesäkuu 2015

---

## Tiivistelmä

Tämän tutkielman päätavoitteena on tutkia trajektorianalyysin soveltuvuutta reaaliaikaisten prosessien tuottaman ison, miljoonia havaintoja sisältävän pitkittäisaineiston analysointiin. Työn alitutkimuskysymyksenä on selvittää trajektorianalyysin mahdollisuuksia tarjota uusia keinoja teleoperaattorin laajakaistaverkon vikojen vähentämiseen.

Lopputyössä tutkitaan mallin polynomiasteen ja ryhmien lukumäärän valintaa sekä aikasarjan pituuden ja otoksen havaintoyksiköiden lukumäärän vaikutusta trajektorianalyysin tuloksiin. Ryhmien lukumäärän määritys on klassinen trajektorianalyysin probleema, mikä nojautuu vahvasti ainakin tämän aineiston kohdalla toimialatietämykseen. Toisaalta, reaalimaailman automaattisten prosessien näkökulmasta ryhmien lukumäärällä ei ole niin merkitystä, kun ne vain palvelevat asetettuja tavoitteita riittävällä tasolla. Ottamalla aineistosta laadultaan ja kooltaan erilaisia otoksia on mahdollista saada hyvin monenlaisia tuloksia. Havaintoyksiköiden lukumäärä ja valittu aikajänne osoittautuvat ratkaiseviksi tulosten käyttökelpoisuuden muodostumisessa.

Vaikka lukuisten trajektorianalyysien jälkeen tutkittavalle aineistolle ei löydy mallia, jota voisi sellaisenaan hyödyntää tuotantotarkoituksiin, analyysien tuloksena saadaan lisätietoa laajakaistaverkon liittymien reaaliaikaisesta toiminnasta ja vikaantumisiin johtavista kehityskuluista. Pääsyy hyvän ennakoinnissa hyödynnettävän mallin puuttumiselle lienee siinä, että tiketöitymiseen johtavat syyt eivät vaikuta riittävällä tasolla ryhmittymiseen ainakaan mallinnukseen valittujen neljän muuttujan avulla. Kuitenkin, kun trajektorianalyysillä tarkastellaan pelkästään tiketillisiä tai tiketittömiä liittymiä, huomataan niiden kehityskuluissa selkeitä eroavuuksia. Erityisesti korostuu myötäsuunnan vaimennuksen rooli katkaistuilla aineistoilla; tiketittömillä kaikkien ryhmien myötäsuunnan vaimennukset ovat vakioarvoiset, kun tiketillisillä mikään niistä ei ole vakio.

Laajakaistaliittymien reaaliprosessit ovat kompleksisia syy- ja seuraussuhteineen, eivätkä yksinkertaiset yksittäiseen arvoon liittyvät kriittisen rajan tarkastelut tai peräkkäisten tapahtumien perusteella tehtävät päättelyt ole osoittautuneet riittävän tarkoiksi. Vianhallinnan laadunkehittäminen on jatkuvaa työtä ja aika näyttää, onko trajektorianalyysillä siihen muuta annettavaa kuin tietämyksen lisääjän rooli. Lopputyön tulosten perusteella menetelmän käytettävyys sovellusalueella on haasteellinen jo senkin takia, että kohteena olevaa laajaa aineistoa voidaan tarkastella vain paloittain. Vaikka sopiva malli löytyisikin jollakin osa-alueella, on epäselvää, toimisiko se yleisemminkin kyseisellä teknologialla, jotta automatisointia voisi lopulta hyödyntää. Lopputyön tulosten perusteella voi kuitenkin todeta, että trajektorianalyysillä on potentiaalia reaaliaikaisten prosessien mallintamisessa.

Asiasanat: trajektorianalyysi, aikasarja, pitkittäisaineisto, mallintaminen, R, flexmix

## Sisällysluettelo

# 1 Johdanto

Pitkittäisaineistolle tehty trajektorianalyysi tarjoaa yleisemmän näkökulman aineistossa esiintyviin erilaisiin latentteihin käyttäytymisprofiileihin sen sijaan, että tarkkailtaisiin kunkin havainnointiyksikön yksilöllistä käyttäytymistä ajassa. Menetelmä on eksploratiivinen, joten sen avulla on mahdollista löytää aineistosta aiemmin havaitsemattomia seikkoja.

Trajektorianalyysiä on käytetty laajalti erityisesti käyttäytymis- ja terveystieteen sekä sosiologian tutkimuksessa viimeisen viidentoista vuoden aikana. Esimerkkejä edellisistä ovat lukuisat alkoholinkäyttötutkimukset (Hill et al 2000, Warner 2005, Berg et al. 2013), sairauden etenemistutkimukset (Dodge et al. 2008) ja hoitovastetutkimukset (Gildengers et al. 2005), työhön kiinnittymisen tutkimus (Virtanen et al. 2010) sekä erilaiset lasten kasvumittareiden ja häiriökäyttäytymisen kehittymistutkimukset (Nummi et al. 2013, Higgins et al. 2013). Näiden lisäksi markkinointitutkimus on hyödyntänyt runsaasti trajektorianalyysia muun muassa asiakassegmentointiin. Muut käyttöalueet ovat selvästi harvemmin edustettuja ainakin julkisten tutkimuspapereiden perusteella, vaikka trajektorianalyysin ajatusmalli onkin yhteensopiva huomattavasti laajemmin toimialasta riippumatta. Muun muassa sosiaalisen verkon dynamiikkatutkimus (Hasan 2012) ja ympäristöön liittyvä tutkimus (Matthews 2014) ovat harvinaisia trajektorianalyysiä hyödyntäviä tutkimuskohteita.

Perussyy nykytilanteen vinouteen lienee se, että menetelmän pääkehittäjät ovat toimineet pääosin edellä mainituilla tyypillisillä toimialueilla, eikä menetelmää ja sen käyttökelpoisuutta vielä tunnisteta laajemmin. Laskennallisten tilastotieteen ja koneoppimisen menetelmien osaaminen ja soveltaminen lienee muutoinkin vahvasti toimialasidonnaista. Laskentakapasiteetin kasvun ja kehittyneiden ohjelmistotyökalujen myötä isonkin aineiston syvällisempi käsittely on taloudellisesti järkevää.

Tämän lopputyön päätavoitteena on tutkia trajektorianalyysin soveltuvuutta reaaliaikaisten prosessien tuottaman ison, miljoonia havaintoja sisältävän aineiston analysointiin. Tutkimusaineisto sisältää erään teleoperaattorin laajakaistaverkon osa-alueen liittymäkohtaista tietoa. Työn alitutkimuskysymyksenä on selvittää, tarjoaako trajektorianalyysi uusia keinoja laajakaistaverkon vikojen vähentämiseen.

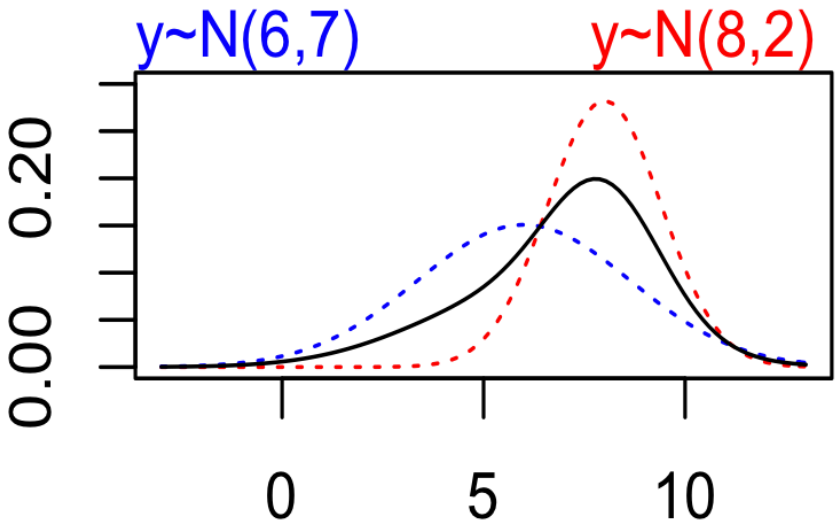
Lopputyön rakenne on seuraava: Luvussa kaksi kuvataan lyhyesti käytetyt menetelmät, niihin liittyvät erityisen mielenkiinnon kohteet ja lopputyössä tehtyjä rajauksia. Tämän jälkeen kuvataan työssä käytettävää aineistoa ja sen puuttuneisuutta. Luvussa neljä esitellään soveltavan aineiston analyysejä ja niissä tehtyjä havaintoja. Lopuksi tehdään yhteenveto lopputyön keskeisimmistä tuloksista.

# 2 Analyysimenetelmät

Pitkittäisaineiston kehityksen mallinnukseen käyttäen lineaarisia malleja on tarjolla useamman tyyppisiä tapoja riippuen käsiteltävänä olevasta aineistosta ja tutkimukseen valitusta näkökulmasta. Pitkittäisaineiston kasvukäyrämalli (Growth Curve Modeling), kasvusekoitemalli (Growth Mixture Modeling) ja trajektorimalli (Group-based Trajectory Modeling) tarkastelevat aineistoa eri lähtökohdista suurimpana erona se, kuinka kukin malli käsittelee yksilöllisiä eroja. (Nagin & Odgers 2010)

Perinteinen kasvukäyrämalli kuvaa kehityksen keskimääräistä kulkua sekä yksilöllistä vaihtelua satunnaisvirheiden varianssin ja kovarianssin avulla. Malli olettaa, että kaikki yksilöt ovat samasta populaatiosta, joten niitä voidaan kuvata yhdellä kehityskäyrällä. Mikäli yksi kehityskäyrä ei riitä kuvaamaan koko populaatiota, tarvitaan ryhmiin pohjautuvia malleja. Kasvusekoitemallissa kukin erillinen osapopulaatio kuvataan omalla kasvukäyrämallilla. (Muthén 1989, 2004)

Kasvusekoite- ja trajektorimallit perustuvat äärelliseen sekoitemalliin (*Finite Mixture Model*), jonka mukaan vastejakauma muodostuu rajallisesta lukumäärästä erilaisia jakaumia, joita ei tunneta etukäteen, vaan ne määräytyvät aineiston perusteella (Nagin & Odgers 2010). Seuraava kuvio 2.1 esittää kahden normaalijakauman sekoitejakaumaa, jossa molempia jakaumia on sekoitteessa samassa suhteessa (50%).




**Kuvio 2.1.** Kahden normaalijakauman sekoitejakauma (musta viiva).

Äärellisessä sekoitemallissa muuttujan **y** tiheysfunktio $f$ esitetään $K$ osajakauman painotettuna tiheysfunktiona

(2.1) $$f(\boldsymbol{y}) = \pi_1 f_1 + \cdots + \pi_K f_K,$$

missä jakaumien sekoitesuhteet $\pi_k$ toteuttavat ehdot $0 \leq \pi_k \leq 1$ ja $\pi_1 + \ldots + \pi_K = 1$ ja $f_\mathrm{k}$ on sekoitejakauman osajakauma. Tyypillisesti aineiston osapopulaatioita ei tunneta etukäteen, joten äärellisen sekoitemallin (2.1) ryhmien lukumäärä, esiintymissuhteet ja jakaumat on estimoitava. (McLachlan & Peel 2000)

Vaikka äärellisen sekoitemallin teoria on yli 100 vuotta vanhaa, aikamuuttujan sisällyttäminen äärelliseen sekoitemalliin on selvästi uudempaa alkaen 1990-luvun lopulta. Tämä lisäys on samalla mahdollistanut uudenlaisen trajektorimallin, jonka sisältämät kehityskäyrät kuvaavat ryhmien käyttäytymistä ajan tai muun ajallisesti etenevän mittarin suhteen. Trajektorimallin pitkittäisaineisto jakautuu useampaan ryhmään, joiden sisäinen käyttäytymisen vaihtelevuus mallin muuttujien suhteen on pienempi kuin ryhmien keskinäinen vaihtelu. Kunkin havainnointiyksikön $\boldsymbol{y}_i$ posteriori-todennäköisyys kuulua ryhmään $\boldsymbol{g}_\mathrm{k}$ lasketaan kaavalla

(2.2) $$p_{\mathrm{ik}} = \frac{\pi_\mathrm{k} f_\mathrm{k}(\boldsymbol{y}_\mathrm{i})}{f(\boldsymbol{y}_\mathrm{i})}, \qquad 0 \leq \pi_K \leq 1, \quad \sum \pi_K = 1.$$

(McLachlan & Peel 2000)

Tämän jälkeen kukin havainto voidaan sijoittaa yhteen ryhmistä – tyypillisesti siihen, jonka posteriori-todennäköisyys on suurin (*hard assignment*). Ryhmän valintaan voidaan käyttää myös esimerkiksi satunnaisvalintaa (*random assignment*) etukäteen valitulla sekoitesuhteella. (McLachlan & Peel 2000, Leisch 2003.) Koska posteriori-todennäköisyys voi ylipainottaa suurimpia ryhmiä (Nagin 2005), suurimman posteriori-todennäköisyyden ollessa alhainen (esimerkiksi 0.4 ja sen alle arvot) satunnaismenettely lienee suositeltavampi, koska silloin arpa ratkaisee kohderyhmän ja vältytään systemaattisilta ylipainotuksilta.

Trajektorimallilla voidaan sopivalla jakaumasekoitteella approksimoida kiinnostuksen kohteena olevan muuttujan tai muuttujien jakaumat. Mitä useampi komponenttijakauma sisällytetään malliin, sitä tarkempi estimaatti saadaan tuloksena. Trajektorimallissa kukin trajektori kuvaa tietyn osapopulaation yksilöllistä kehityskulkua. Populaation vaihtelua ryhmien välillä mallinnetaan täten trajektoreiden lukumäärän ja muodon kautta. (Nagin 2005)

Suurin ero edellisten ryhmiin perustuvien mallien välillä on se, että kasvusekoitemallissa ryhmän kehityskulku sisältää satunnaisvaihtelun ja trajektorimallissa ei (Nagin & Odgers 2010). Tässä työssä rajaudutaan jatkossa käsittelemään Naginin (2005) trajektorimallia. Työn päälähteet ovat Nagin 2005 yhdessä McLachlan & Peel 2000 kanssa.

## 2.1 Trajektorianalyysi

Trajektorianalyysi tarjoaa yleisemmän näkökulman aineistossa esiintyviin erilaisiin käyttäytymisprofiileihin sen sijaan, että tarkkailtaisiin kunkin havainnointiyksikön yksilöllistä käyttäytymistä yksi kerrallaan. Tämä tarjoaa mahdollisuuden keskittyä tarkastelemaan esimerkiksi normaalista poikkeavien patologisten tapausten käyttäytymistä ryhmänä ja mahdollisesti estää vahingollinen kehityskulku minimoiden kokonaisvaikutuksia. (Nagin 2005, Nagin & Odgers 2010)

Trajektorianalyysin tavoitteena on löytää aineistosta ennalta tuntemattomia ryhmiä, joiden sisältämät yksilöt käyttäytyvät samantapaisesti. Tyypillisesti löydetään eri tasoilla toimivia ryhmiä, joiden kehityskulku on nouseva, laskeva tai tasainen. Koska ryhmitys perustuu todennäköisyyksiin, ryhmien kokoonpano ei ole kiinteä eivätkä klassiset tilastolliset päättelymenetelmät ole käyttökelpoisia. On siten hyvä muistaa, että parhaimmillaankin saatu kokoelma ryhmiä tarjoaa vain approksimaation jatkuvan jakauman kuvaamiseen. Trajektorianalyysi on semiparametrinen menetelmä, jossa ryhmien lukumäärä ja sisältö estimoidaan parametrittomasti, mutta trajektorimallien polynomikertoimet estimoidaan parametrisilla menetelmillä. (Nagin 1999, 2005)

Oletetaan, että $\boldsymbol{y_i} = (y_{i1} \dots y_{it})'$ on ajassa t havaittujen *i*:nnen mittausyksikön muuttujan y arvoja. Tällöin havaintojen $\boldsymbol{y}_i$ reunatodennäköisyysfunktio $f(\mathbf{y}_\mathrm{i} \mid \mathbf{X}_i, \boldsymbol{z}_\mathbf{i})$ voidaan määrittää K:n ryhmän sekoitteena kaavalla

$$f(\mathbf{y}_\mathrm{i} \mid \boldsymbol{X}_i, \boldsymbol{z}_\mathbf{i}) = \sum_{k=1}^{K} \pi_\mathrm{k}(\boldsymbol{z}_\mathbf{i}) f_\mathrm{k}(\mathbf{y}_\mathrm{i} \mid \boldsymbol{X}_i) \ , \quad \sum \pi_\mathrm{k} = 1, \ \pi_\mathrm{k} \geq 0, \tag{2.3}$$

missä vektori $\boldsymbol{X}_i$ sisältää kaikki kovariaatit ja $\boldsymbol{z}_i$ kaikki ajasta riippumattomat kovariaatit. Edellä esitetyn yhtälön $\pi_\mathrm{k}(\boldsymbol{z}_\mathbf{i})$ on tutkittavan havainnon todennäköisyys kuulua ryhmään k ja $f_\mathrm{k}(\mathbf{y}_\mathrm{i} \mid \boldsymbol{X}_i)$ on k:nnen ryhmän tiheysfunktio. (McLachlan & Peel 2000)

Trajektorianalyysi perustuu suurimman uskottavuuden menetelmään, joka on riippuvainen valituista todennäköisyysjakaumista. Eksponentiaaliseen jakaumaperheeseen, esimerkiksi normaali-, binomi-, Poisson- ja gammajakaumien yhdistelmä tuottaa täten yleistettyjen lineaaristen mallien sekoitejakauman. Edellisen erikoistapauksena normaalijakaumien yhdistelmä vastaa regressiomallien sekoitejakaumaa. Tuntemattomat muuttujien estimaatit $\theta_i$ saadaan maksimoimalla M havainnon $\boldsymbol{y}_1 \dots \boldsymbol{y}_M$ logaritmoitu uskottavuusfunktio

$$logL(\boldsymbol{\theta} \mid \boldsymbol{y}_1 \dots \boldsymbol{y}_M) = \sum_{i=1}^{M} \log f(\boldsymbol{y}_i \mid \boldsymbol{X_i}, \boldsymbol{z}_i). \tag{2.4}$$

(McLachlan & Peel 2000)

### 2.1.1 Suurimman uskottavuuden estimointi EM-algoritmillä

Suurimman uskottavuuden (*Maximum likelihood, ML*) estimointiin käytetään numeerisia optimointialgoritmeja (Dempster et al.1977), joista esimerkkinä eniten käytetty iteratiivinen EM-algoritmi, joka sisältää kaksi päävaihetta: E-vaiheessa (*expectation*) mallin 'puuttuvat arvot' korvataan estimaateilla ja M-vaiheessa (*maximization*) maksimoidaan E-vaiheessa saatu tulos. Näitä kahta vaihetta toistetaan perätysten, kunnes laskenta suppenee (*converge*) eikä tulos enää tästä tarkennu merkitsevästi, jotta iterointia kannattaisi enää jatkaa.

Kukin estimaattiyhdistelmä tuottaa erisuuruisen log-uskottavuuden arvon, joka kasvaa sitä mukaa, kun iteraatiot etenevät. Tavoitteena on löytää estimaatit (ryhmittäiset $\beta$-kertoimet, varianssit ja havainnot), jotka ovat suurimmalla todennäköisyydellä tuottaneet aineiston, toisin sanoen, niillä estimaateilla täydennetty malli sovittuu parhaiten tarkasteltavaan aineistoon. Iteroinnin päättyminen voidaan määrittää asettamalla minimiraja log-uskottavuuden arvon muutokselle tai maksimi iteraatioiden lukumäärälle. (Leisch 2003, Grün & Leisch 2008, Enders 2010)

E-vaiheessa lasketaan havaintoyksikön posteriori-todennäköisyydet kullekin ryhmälle *k* mallin sisältämien estimoitujen muuttujien $\theta_k$ avulla

$$\hat{p}_{\mathrm{nk}} = \frac{\pi_{\mathrm{k}} f_{\mathrm{k}}(\boldsymbol{y}_{\mathbf{n}}|\boldsymbol{X}_{\boldsymbol{n}},\boldsymbol{\theta}_{\boldsymbol{k}})}{\sum_k \pi_k f(\boldsymbol{y}_{\mathbf{n}}|\boldsymbol{X}_{\boldsymbol{n}},\boldsymbol{\theta}_{\boldsymbol{k}})}, \tag{2.5}$$

minkä jälkeen saadaan estimaatit ryhmien priori-todennäköisyyksille

$$\hat{\pi}_{\mathrm{k}} = \frac{1}{N}\sum_{n=1}^{N} \hat{p}_{nk}. \tag{2.6}$$

Posteriori-todennäköisyyksiä käytetään painoarvoina M-vaiheen log-uskottavuuden maksimoinnissa

$$max_{\theta_k} \sum_{n=1}^{N} \hat{p}_{\mathrm{nk}} log\, f_k(y_n|x_n, \theta_k). \tag{2.7}$$

(Leisch 2003)

Lopputyössä käytetyssä R:n *flexmix*-paketissa tehdään automaattista ryhmien lukumäärän määritystä aina ennen M-vaihetta. Tällöin redundanteiksi, toisin sanoen liian pieniksi, osoittautuneet ryhmät hylätään mallista ja lopputulos voi täten sisältää vähemmän ryhmiä alkuperäiseen pyyntöön verrattuna. Lisäksi E-vaiheen jälkeen huolehditaan tarvittaessa siitä, että havaintoyksiköiden kaikki havainnot kuuluvat aina vain yhteen ryhmään, jota tietoa hyödynnetään seuraavassa M-vaiheessa. (Grün & Leisch 2008)

EM-algoritmin tyypillisinä haasteina ovat suppenemisen hitaus ja mahdollinen paikallisen maksimin löytyminen globaalin maksimin sijaan. Numeeriset laskennalliset ongelmat tulevat myös helposti vastaan estimoinnin aikana, erityisesti silloin, kun käsitellään hyvin lähellä nollaa olevia arvoja. Lokaalin maksimin välttäminen onnistuu parhaiten tekemällä suurimman uskottavuuden

estimointi useampia kertoja eri alkuarvoin, minkä jälkeen valitaan vaihtoehto, joka maksimoi uskottavuuden. (Leisch 2003, Eldén et al. 2004, Grün & Leisch 2008)

### 2.1.2 Toistomittausaineisto

Mikäli aineisto sisältää toistomittauksia, joissa *M* havaintoyksiköllä on yhteensä *N* havaintoa, yhtälö (2.4) saa muodon

$$logL = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \log f(\boldsymbol{y}_{mn} \mid \boldsymbol{X}_{mn}, \boldsymbol{z}_{mn}), \quad \sum_{m=1}^{M} N_m = N, \tag{2.8}$$

missä $(y_{mn}, x_{mn})$ on *m*:nnen havaintoyksikön *n*:s havainto. Posteriori-todennäköisyys havaintoyksikölle *m* kuulua ryhmään *j* on

$$\mathbf{P}(j \mid m) = \frac{\pi_j \prod_{n=1}^{N_m} f(\boldsymbol{y}_{mn} \mid \boldsymbol{x}_{mn,\theta_j})}{\sum_k \pi_k \prod_{n=1}^{N_m} f(y_{mn} \mid \boldsymbol{x}_{mn}, \theta_k)}. \tag{2.9}$$

(Leisch 2003)

### 2.1.3 Mallin muodon valinta

Trajektorianalyysin mallin valintaan vaikuttaa monia asioita. Yksi keskeinen tekijä on aineistosta valitun vastejakauman muoto. Usein tavoitellaan normaalijakaumaa ja tarvittaessa käytetään erilaisia muunnoksia, jos vaste ei ole jo valmiiksi riittävän normaalinen. Box-Cox muunnos (Box & Cox 1964) on useimmiten käytetty menetelmä näihin tarkoituksiin. Osa trajektorianalyysin toteutuksista (kuten lopputyössä käytetty R:n kirjasto *flexmix*) tukee lisäksi yleistettyjä lineaarisia malleja, jolloin vaste voi olla muodoltaan esimerkiksi Poisson tai gamma, eikä muunnoksia tarvita. Vaikka muunnosten käyttö olisikin suoraviivaista, tulosten tulkinta vaikeutuu, joten muunnosten teossa on syytä käyttää tarkkaa harkintaa ja vaihtoehtoisesti etsiä uudenlaisia työkaluja ja menetelmiä, joilla muunnokset voisi minimoida. Tyypillisimpiä käytettyjä muunnoksia ovat logaritmi-, neliöjuuri- ja käänteismuunnokset sekä potenssiin korottaminen.

Mallin aikaa kuvaavan selittäjän polynomiasteen valinta määrittää ryhmän vastetrajektorin muodon. Esimerkiksi kolmannen asteen polynomi kuvaa vasteen $y_{it}^*$ kolmannen asteen riippuvuutta iästä mallilla

$$y_{it}^* = \beta_0^j + \beta_1^j ik\ddot{a}_{it} + \beta_2^j ik\ddot{a}_{it}^2 + \beta_3^j ik\ddot{a}_{it}^3 + \varepsilon_{it}. \tag{2.10}$$

Yhtälössä $ik\ddot{a}_{it}$ on yksilön i ikä ajankohtana t ja $\varepsilon_{it}$ on mallin normaalijakaumaa noudattava virhetermi, jonka odotusarvo on nolla ja varianssi $\delta^2$. Jokaisella trajektorilla on oma malli ja sen $\beta^j$-kertoimet, missä *j* identifioi tarkasteltavana olevan trajektorin. Vastaavasti toisen asteen polynomi määrittää neliöllisen trajektorin, jossa $\beta_3$ on asetettu nollaksi. Ensimmäisen asteen polynomi vastaa li-

neaarista suoraa, jossa $\beta_2$ ja $\beta_3$ ovat nollia. Nollan asteen polynomi taasen on vaakasuora viiva, jonka $\beta_0$-kerroin määrää tasokorkeuden muiden kerrointen ollessa nolla. (Nagin 2005.) Käytännössä aineiston perusteellinen estimointi ei kuitenkaan aina onnistu ja tällöin tavoitteeksi jää lähinnä mallioletuksen maksimointi.

Yhden vasteen mallin sijaan voidaan rakentaa kahden (*dual trajectory model*) tai useamman (*multi-trajectory model*) vasteen malli, joissa voidaan tarkastella valitun vastedimension trajektoreiden samanaikaista taikka kokonaan tai osittain peräkkäistä kehitystä. Kaksi- ja monitrajektorianalyysin avulla saadaan tietoa mallissa tutkittavana olevien vasteiden erillisistä mutta samalla toisiinsa liittyvistä vaikutuksista. (Nagin & Tremblay 2001, Nagin & Odgers 2010)

Mallin muodon valintaan vaikuttaa merkitsevästi aineiston havaintoyksiköiden ja ajallisten mittapisteiden lukumäärä. Aiemmissa tutkimuksissa on havaittu, että 300 - 500 havaintoyksikön ja niitä suuremmissa saman sovellusalueen analyyseissä trajektoriryhmien lukumäärä säilyy samana. Aikamittauspisteiden lukumäärää selvästi muuntamalla, esimerkiksi kahdentamalla aikasarjan pituus, voi olla merkittäviä vaikutuksia saatuihin tuloksiin. (D'Unger et al. 1998, Sampson et al. 2004)

### 2.1.4 Ryhmien lukumäärän valinta

Ryhmien lukumäärän määritys on klassinen trajektorimallin valintaongelma, eikä tähän ole edelleenkään saatavilla tyhjentävää vastausta. Trajektorianalyysillä tavoitellaan ryhmäkohtaisia malleja siten, että jokaisella sekoitejakauman sisältämällä osajakaumalla on oma mallinsa. Ryhmien lukumäärä voi vapaasti määräytyessään olla joko tarpeettoman suuri tai hyödyttömän alhainen ja kaikkea tältä väliltä. Estimoinnin suppeneminen voi aiheuttaa ongelmia, jos tavoitellaan tarpeettoman suurta määrää tai vastaavasti suppenemisen tuloksena saadaan hyödyttömän vähän ryhmiä. Kullekin ryhmälle on aina hyvä löytää joku sovellusalueen tietämyksen pohjalta koottu merkitys, ja jos sitä ei löydy, tulee harkita ryhmien yhdistämistä pienentämällä tavoiteltua ryhmien lukumäärää. Riittävän luotettavan sovellusalueen selityksen puuttuessa on hyvä tiedostaa todellinen riski sille, että trajektorianalyysin avulla aineistosta on mahdollista löytää myös keinotekoisia trajektoreita, jos mallin sovitus on puutteellinen.

Koska ryhmään kuuluvien havaintojen joukko ei ole kiinteä ja ryhmään kuuluminen on siten aina jollain tasolla epävarmaa, klassisia tilastollisia testejä, kuten F-testi tai $\chi^2$-testaus, ei voida käyttää testaamaan ryhmien välisten erojen merkitsevyyttä (Roeder et al. 1998). Mallien välisiä eroja voidaan kuitenkin tarkastella Bayesin informaatiokriteerin (*Bayesian Information Criterion, BIC*) avulla. BIC:ä voidaan käyttää laajalti erilaisten mallien keskinäiseen vertailuun oletuksella, että mallien priori-todennäköisyydet ovat yhtäsuuret. Malli, jonka posteriori-todennäköisyys on suurin, minimoi BIC:n, joten malli, jolla on pienin BIC on suositeltavin. On kuitenkin hyvä tiedostaa, että koska saatu BIC-arvo on aina approksimaatio, BIC-tulosta voidaan käyttää parhaimmillaankin vain yhtenä mallinvalinnan aputyökaluna. Likimääräinen BIC lasketaan kaavalla

(2.11) $$BIC = -2\log(L) + k*\log(N),$$

missä L on mallin uskottavuusfunktio ja N on otoskoko. Yhtälön k on mallissa olevien parametrien lukumäärä, mikä huomioi mallin polynomiasteen ja ryhmien lukumäärän. Esimerkiksi kolmen ryhmän toisen asteen polynomimallin parametrien lukumäärä on 3*(3+1), toisin sanoen, ryhmien lukumäärä kertaa regressiomallin β-kerrointen lukumäärä lisättynä varianssilla. (Schwarz 1978, Kass & Raftery 1995, Nagin 2005)

### 2.1.5 Puuttuva tieto aineistossa

Trajektorianalyysi ei vaadi tasapainoista (*balanced*) pitkittäisaineistoa, jossa kaikilla tutkimusyksiköillä on kaikkien muuttujien arvot aina samoissa ajallisissa mittauspisteissä. Pitkittäisaineistoille onkin hyvin tyypillistä, että puuttuneisuutta esiintyy. Aineiston puuttuneisuus voi aiheuttaa ongelmia erityisesti, jos arvoja puuttuu runsaasti joko paikallisesti tai yleisesti. Hyväksyttävän puuttuneisuuden taso on sovellusalueesta ja puuttuneisuuden rakenteesta riippuvainen. Yleistäen voidaan kuitenkin todeta, että mitä eheämpi pitkittäisaineisto on, sitä luotettavampia tuloksia saadaan. (Nagin 2005)

Suurimman uskottavuuden menetelmät toimivat luotettavasti sekä MCAR- (*Missing completely at random*) että MAR-tapauksissa (*Missing at random*). MCAR-tyyppisessä aineistossa puuttuneisuus on täysin satunnaista eikä riipu täten siitä, mitä on havaittu tai oltaisiin voitu havaita. MAR-tyyppisessä aineistossa puuttuneisuus voi riippua siitä, mitä on havaittu, mutta ei voi riippua siitä, mitä olisi voitu havaita. MAR-tyypisen puuttuneisuuden rakenne lienee yleisin käytännön sovellutuksissa. MNCAR-tyyppisen (*Not missing at random*) puuttuneisuuden rakenne on jo haasteellisempi ja suurimman uskottavuuden menetelmän tulosten harhaisuus (*bias*) kasvaa merkittävästi ainakin jossain osassa saatua mallia. (Enders 2010)

Suurimman uskottavuuden menetelmä käsittelee puuttuvaa tietoa satunnaisena ilmiönä, joten käytössä oleva aineisto voidaan hyödyntää maksimaalisesti ilman tulosten vääristymistä, mikäli puuttuvan tiedon rakenne on aidosti MAR-tyyppiä ja käytössä oleva malli jakaumaoletuksineen on luotettavalla tasolla. EM-algoritmilla voidaan estimoida kokonaan puuttuvia arvoja, joista esimerkkinä työn aiheeseen keskeisesti liittyvät latentit trajektorit (Grün & Leisch 2008). Käytännön sovellutuksissa puutteellinen havainto voi kuitenkin olla käyttökelvoton ja se joudutaan poistamaan aineistosta ennen analyysiä esimerkiksi silloin, kun käytetty ohjelmistopaketti ei tue riittävällä tasolla puuttuvaa tietoa sisältävän aineiston analysointia. (Enders 2010)

### 2.1.6 Kovariaattien sisällyttäminen malliin

Perustrajektorianalyysissä havainnot oletetaan keskenään riippumattomiksi trajektorin sisällä. Kovariaatteja voidaan kuitenkin sisällyttää analyysiin antamaan lisäselitystä. Aikariippuvaisten kovariaattien avulla voidaan mallintaa erilaisia

trajektorin kehityskulkua potentiaalisesti muuntavia ulkoisia impulsseja. Edellä esitetyn kolmannen asteen polynomimallin (2.10) täydennys kovariaateilla $z_l$ on

$$(2.12)\quad y_{it}^{*} = \beta_0^j + \beta_1^j ik\ddot{a}_{it} + \beta_2^j ik\ddot{a}_{it}^2 + \beta_3^j ik\ddot{a}_{it}^3 + \alpha_1^j z_{1t} + \cdots + \alpha_L^j z_{iLt} + \varepsilon_{it},$$

missä $z_1$ ... $z_L$ ovat kovariaatteja ja $\alpha_1$ ... $\alpha_L$ niiden kertoimia, L määrittää kovariaattien lukumäärän. (Nagin 2005)

Tässä lopputyössä kovariaatteja ei ole sisällytetty trajektorimalleihin, koska tavoitteena on ollut löytää uutta tietoa uudenlaisen aineiston ryhmityksen avulla ilman etukäteen kiinnitettyjen muuttujien, esimerkiksi liittymäprofiilin, vaikutusta lopputulokseen. Vakioisista luokittelutyyppisistä kovariaateista onkin jo varsin paljon tietoa aiemmin tehtyjen sovellusalueen analyysien tuloksena. Lisäksi potentiaalisten aikariippuvien kovariaattien hyödyntäminen ei osoittautunut tärkeäksi lopputyössä käytetyn aineiston kohdalla.

## 2.2 Yleistetyt additiiviset mallit

Trajektorimallin kuvaamiseen käytetään yleistettyjä additiivisia malleja (Generalized additive model, GAM), joiden avulla saadaan piirrettyä kaikkien kokonaismalliin sisältyvien vastemuuttujien trajektorit ryhmittäin. Tämän jälkeen tarkastellaan saatuja trajektoreja ryhmittäin ja haetaan toimialakohtaisia selityksiä kunkin ryhmän merkitykselle.

Yleistetyt additiiviset mallit ovat yleistettyjen lineaaristen mallien (Generalized linear model, GLM) yleistys, jonka avulla monimutkaisten riippuvuuksien mallinnus onnistuu joustavammin, eikä tarvita etukäteen määriteltyjä muuttujamuunnoksia tai polynomiastevalintoja. Yleistetyissä additiivisissa malleissa tarvittavat muodon valintaan liittyvät muunnokset estimoidaan epäparametrisesti mallinnusprosessin aikana käyttäen erilaisia tasoitusfunktioita $f_j$. Malleissa sallitaan myös lineaariset termit, joiden mallintamiseen ei käytetä tasoitusfunktiota tarpeettomasti. Seuraava yhtälö kuvaa additiivisen perusmallin

$$(2.13)\quad y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + Z\gamma + \epsilon,$$

missä Z sisältää mallin regressiomuuttujat (kategoriset tai numeeriset) estimoituine γ-parametreineen. (Faraway 2006)

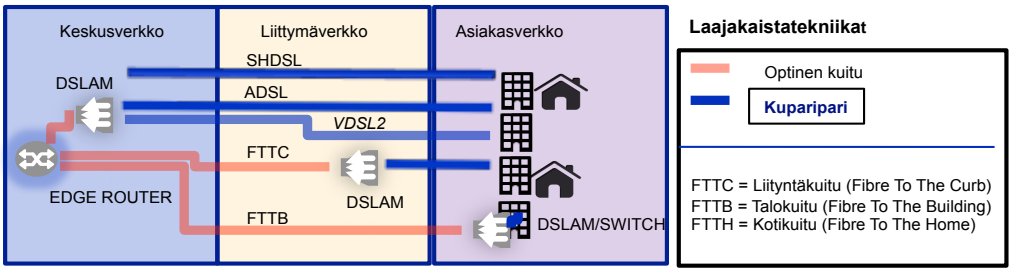
Lopputyössä käytetään R:n *mgcv*-kirjastoa, joka käyttää automaattista tasoitusparametrien valintaa siten, että käyrä kulkee kaikkien havaintopisteiden läheltä, mutta ei välttämättä kautta (*penalized smoothing spline*). Muuttujakohtaisen suodatuksen tasoa säätelee kerroin, joka tuotetaan yleistetyn ristiinvalidoinnin avulla (*generalized cross-validation, GCV*). (Faraway 2006)

# 3 Aineiston kuvaus

Vikojen hallinnalla ja ennakoinnilla on tärkeä rooli tietoliikennepalveluiden kustannusten tehostamisessa ja palvelun laadun parantamisessa. Tavoitteena on korjata mahdollisimman paljon piileviä vikoja ennen kuin asiakas edes huomaa, että on jotain ongelmaa. Tyypillisesti vikaantumisesta aiheutuvat kokonaiskustannukset ovat riippuvaisia ajankohdasta, jolloin vikaantuminen havaitaan. Esimerkiksi kaapeleiden kastumistilanteissa mitä aikaisemmin kastuminen havaitaan ja ongelma voidaan korjata, sitä pienemmillä kokonaiskustannuksilla päästään (Hyppänen 2007). Laadukkaan vianhallinnan avulla tietoliikenneoperaattori pystyy paremmin täyttämään lukuisat viranomaisvaatimukset sekä ylläpitämään tai jopa parantamaan asiakkaiden kokonaistyytyväisyyttä, mikä viime kädessä näkyy toiminnan kannattavuutena.

Lopputyön soveltava osuus liittyy erään tietoliikenneoperaattorin DSL-tekniikalla (*Digital subscriber line*) toteutetun laajakaistaverkon sähköisiin mittauksiin ja käyttäjän avaamiin vikailmoituksiin, joihin tästä lähtien viitataan sanalla tiketti. Sähköisillä mittauksilla saadaan toiminnallista tietoa tietoliikenneoperaattorin liityntäverkosta, joka sisältää optisen runkoverkon rajapinnan ja asiakasliittymän parikaapelitoteutuksen välisen verkko-osuuden. Mittausten avulla on mahdollista saada tilannetietoa siitä, miten mielenkiinnon kohteena oleva tietoliikenneverkon osa toimii. Tätä tietoa tarvitaan muun muassa palvelun laadun kehittämiseen ja asiakaspalvelutarkoituksiin. Mittauksia tehdään säännöllisesti päivittäin.

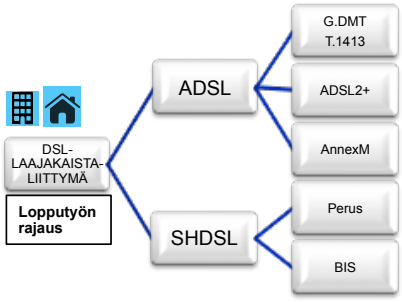
Lopputyön tavoitteena on tutkia trajektorianalyysin mahdollisuuksia auttaa vähentämään ja ennakoimaan vikaantumisia päivittäisten mittausten tuottamaa pitkittäisaineistoa käyttäen. Erillisiä tikettitietoja käytetään trajektoreiden ohella tuomaan lisäsyvyyttä analyysiin. Oheinen kuvio 3.1. esittää sovellusalueen ylemmän tason teknisen arkkitehtuurin siltä osin, mikä on oleellinen lopputyön aineiston näkökulmasta.




**Kuvio 3.1.** Laajakaistaverkon ylemmän tason toteutusarkkitehtuuri.

DSLAM-laite (DSL access multiplexer) kytkee useamman asiakasliittymän nopeayhteyksiseen runkoverkkoon. Aineiston asiakasliittymän tekninen toteutus vaihtelee tarpeen mukaan erityyppisten ADSL- ja SHDSL-modeemien

välillä. Kuvio 3.2. kuvaa tarkemmin aineistossa olevien DSL-liittymien vaihtoehdot käytetyn kättelyprotokollan (*handshake*) mukaan. Kättelyprotokolla kuvaa DSLAM ja xDSL-modeemien välisen neuvottelustandardin, minkä avulla laajakaistapalvelu lopulta toteutetaan. Protokolla määrittelee muun muassa yhteyden välillä käytettävän tietoliikenteen maksiminopeuden.




**Kuvio 3.2.** Aineiston DSL-laajakaistan liittymätyypit.

## 3.1 Muuttujat

Koko pitkittäisaineisto sisältää yhteensä 3938174 mittausta 23136 liittymästä (tilastoyksikkö) kuuden kuukauden ajalta. Kullakin aineistoon sisältyvällä liittymällä on yhteensä 1 - 184 mittausta, koska kyseessä on päivittäismittausaineisto, jolloin liittymällä on korkeintaan yksi mittaus kunakin tarkastelujakson päivänä. Aineisto ei ole täten tasapainoinen, koska mittausten lukumäärä vaihtelee liittymittäin. Aineiston keskeisimmät muuttujat ovat liittymän yksilöivä tunniste, mittauspäivä, myötä- ja paluusuunnan vaimennus ja kohinamarginaali. Aineiston kattava muuttujalista on esitelty liitteessä A.

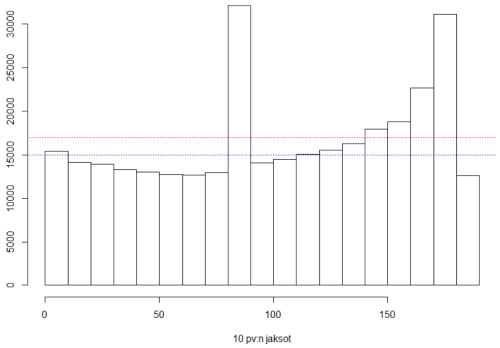
## 3.2 Puuttuva tieto

Reaaliaikaisessa prosessoinnissa on monenlaisia syitä, miksi osa mittauksista puuttuu joko kokonaan tai mittauksen sisältö ei ole täydellinen. Mahdolliset syyt ovat muuan muassa seuraavankaltaisia: tietoliikenneverkon häiriöt estävät mittausten keruun kokonaan vaikutusalueellaan, liittymien mittauskonfiguraatioissa on vaihtelua sekä mitattavien muuttujien että koko mittauksen osalta ja päätelaite on pois verkosta esimerkiksi sähkökatkon tai laitteen poiskytkennän takia kokonaan. Lisäksi verkon laite- ja ohjelmistopäivitysten takia mittausten keruussa voi olla puutteita rajatulla aikavälillä. Seuraavaksi tutkitaan tarkemmin minkälaisesta puuttuneisuudesta lopputyön aineiston kohdalla on kyse ja mitä sen perusteella tehdään ennen varsinaista analyysivaihetta.

### 3.2.1 Mittausten puuttuminen kokonaan

Jos kaikilla liittymillä olisi maksimimäärä mittauksia tarkastelujaksolla, niin mittausten lukumäärä olisi 4257024. Mittausten puuttuneisuus on täten noin 7.5%. Kuviossa 3.3. olevan histogrammin perusteella puuttuneisuus on varsin tasaista tarkastelujaksolla lukuun ottamatta piikkiä lokakuun loppupuolella (mittauksen estänyt tietoliikennevika yhtenä päivänä) ja nousevaa trendiä loppukaudesta (poikkeuksellisesti osa liittymistä vailla mittauksia laajavaikutteisen verkkopäivityksen takia).

Histogrammiin sinisellä merkitty tyypillinen päivittäinen puuttuneisuuden perustaso on noin 1400 mittausta, kun poikkeukselliset tilanteet on jätetty huomioimatta. Arviolta noin 70% tästä perustasosta selittyy sillä, että liittymien lukumäärä vaihtelee tarkastelujaksolla eikä kyse ole varsinaisesti puuttuvista mittauksista vaan siitä, että pitkin aikajanaa uusia liittymiä tulee mukaan ja osa vanhoista poistuu. Matemaattinen punaisella viivalla merkitty keskiarvo on 1733 mittausta. Histogrammin kukin pystypalkki sisältää aina kymmenen päivän puuttuvien mittausten lukumäärän. Viimeinen palkki sisältää poikkeuksellisesti vain neljän päivän puuttuvat mittaukset, joten se ei ole suoraan vertailukelpoinen muiden histogrammin osien kanssa.




**Kuvio 3.3.** Puuttuvat mittaukset aikajaksolla 1.8.2014-31.1.2015.

Puuttuviin mittauksiin liittyviä käyttäjän avaamia tikettejä on yhteensä 64 kappaletta ja onnistuneisiin mittauksiin liittyviä tikettejä on 962. Puuttuviin mittauksiin (0.020%) liittyy täten suhteessa hieman vähemmän käyttäjän avaamia tikettejä kuin onnistuneisiin mittauksiin (0.024%), joten mittausten puuttuneisuus ei näyttäisi olevan riippuvainen tikettien esiintyvyydestä. Puuttuneisuus noudattanee täten joko MCAR- tai MAR-rakennetta, jotka molemmat ovat toimivia trajektorianalyysin näkökulmasta.

Toisaalta, kun huomioi sen, että noin 70% puuttuneisuudesta johtuukin liittymien poistumisesta ja uusien tulosta pitkin aikajanaa, tikettien osuus aidosti puuttuvilla mittauksilla on noin 0.067%. Tämä on noin kolme kertaa onnistuneiden mittausten osuus, mikä herättää epäilystä siitä, että aiempi oletus MCAR- tai MAR-rakenteesta ei olekaan täysin pitävä. Lisävarmistusta saadakseen tikettejä pitäisi tarkastella vielä lisää ja eliminoida mahdollisesti samasta tilanteesta saa-

tujen useampien tikettien vaikutus. Kun suhteellisen osuuden suuruusluokka kuitenkin pysyy edelleen pienenä, jatketaan MCAR/MAR oletuksella.

### 3.2.2 Mittauksen muuttujien arvojen osittainen puuttuminen

Mittausten muuttujien arvojen puuttuneisuus vaihtelee nollasta 19%:iin. Suurimmat puutteet liittyvät lähetystehoihin siten, että tietyn tyyppiset liittymät ovat yliedustettuina. Muiden muuttujien puuttuneisuus on hyvin pientä: vaimennus 3 - 6% ja kohinamarginaali 0.07 - 0.2%. Liitteessä B on kuvattu muuttujien tarkempi puuttuneisuuden rakenne.

Lähetysteho korreloi voimakkaasti vaimennuksen kanssa. Lisäksi aiempien analyysien ja toimialapohjaisen tietämyksen perusteella on oletettavaa, että lähetystehon tuoma lisäinformaatio on vain marginaalista tulosten näkökulmasta, jos vaimennus on sisällytetty analyysiin. Koska käytössä oleva R:n työkalu ei hyväksy puutteellisia mittausrivejä, jätetään täten suurimman puuttuneisuuden omaavat lähetystehot kokonaan pois trajektorianalyysin käyttämästä aineistosta. Tällöin vaimennuksen ja kohinamarginaalin tuottama puuttuneisuus on enää 7.2% (yhteensä 283794 mittausta) alkuperäisestä mittausaineistosta maksimoiden käytössä olevan keskeisen informaation. Alkuperäisen aineiston, joka sisältää myös puutteelliset havainnot, ja vain täydellisten havaintojen sisältämän aineiston tunnusluvut (liite B) poikkeavat osittain toisistaan. Täten puuttuneisuuden rakenne varmistuu MAR-tyyppiseksi, kun hylätään pieni epävarmuus kokonaan puuttuvien mittausten osalta (luku 3.2.1).

### 3.2.3 Raaka-datan laadullisia ominaisuuksia

Aineiston käytettävyys, analytiikan toteutus ja tulokset ovat aina enemmän tai vähemmän riippuvaiset useammasta lähteestä kerättävän datan muodosta ja käytettävyydessä. Formaalisti muotoiltu tieto on laadukkaan analytiikan perusvaatimus. Lisäksi tiedon keruussa käyttäjälle tarjottavat vaihtoehdot on syytä asettaa minimiin, jotta analyyseissä päästäisiin hyödyllisempiin koonteihin. Pääperiaate on kerätä tietoa vain sillä tarkkuudella, mitä myös käytetään päätöksentekoon tavalla tai toisella. Liitteessä C on esimerkkejä datankeruuseen liittyvistä huomionarvoisista periaatteista, joihin lopputyönkin osalta tuli käytännön näkökulmaa.

# 4 Aineiston analyysi

Lopputyön päätutkimuskysymyksenä on tutkia trajektorianalyysin käytettävyyttä, kun tarkasteltavana on iso reaaliaikaisen järjestelmän tuottama aineisto. Alitutkimuskysymyksenä on saada tarkempaa tietoa trajektorianalyysin mahdollisista hyödyistä laajakaistaverkon liittymien häiriönhallinnassa.

Liittymien vikojen taustasyynä ovat usein tilapäiset, hetkelliset tietoliikenteen vaihteluun liittyvät häiriöt, jotka xDSL-modeemi pystyy itsekin normaalisti korjaamaan automaattisesti niin, että loppukäyttäjä ei edes huomaa häiriötä. Jos kuitenkin tiketti on luotu ja asiakkaan vikaa aletaan selvittämään, tilanne on usein jo ohi eikä ole mitään tehtävää. Pitkäkestoisimmissa ongelmissa taasen ongelmien aiheuttajana voivat olla erilaiset liittymän konfiguraatiovirheet, kaapeliyhteyteen liittyvät tekijät, sähkönsyöttöongelmat tai laiteviat, joihin vaaditaan tyypillisesti jonkintasoinen korjaustoimenpide. Oletus on, että osaa näistä pitkäaikaisista vioista on mahdollista myös ennakoida ja sitä kautta minimoida vikaantumisesta aiheutuvia kokonaiskustannuksia, mistä esimerkkinä on kaapeleiden kastuminen.

Nykyinen automaattinen vianennakointi perustuu pääosin sähköisten mittausten kriittisten tasojen ylityksiin tai alituksiin sekä verkosta kerättyjen tapahtumien (*event*) peräkkäisten esiintymisten seuraamiseen. Käytännössä on kuitenkin havaittu, että nämä tavat synnyttävät tarpeettoman paljon turhia ja työllistäviä vikailmoituksia, minkä takia uudenlaisten menetelmien tutkiminen ja käyttöönotto on kiinnostavaa. Uusien havaintojen ja tietämyksen kasvun kautta automatiikkaa voidaan kehittää muun muassa parantamalla olemassa olevan ennakoinnin tarkkuutta sekä löytämällä uudenlaisia häiriövaikutuksia ja niitä indikoivia herätteitä.

Päivittäisillä liittymän sähköisten arvojen mittauksilla saadaan ajankohtaista tietoa liittymän reaaliaikaisesta toiminnasta. Lopputyön alitavoitteena on tutkia trajektorianalyysin mahdollisuuksia havaita ongelmallisia liittymän kehityskulkuja, joihin voitaisiin puuttua parhaassa tapauksessa jo ennakoivasti ennen kuin asiakas on edes havainnut ongelmia. Mikäli hyvä trajektorimalli löytyy, sitä voisi mahdollisesti käyttää ennakoivasti seuraamalla liittymien ryhmittymistä ja ajassa tapahtuvia siirtymiä ryhmien välillä. Koska laajakaistaverkon liittymien kokonaislukumäärä on suuri, ongelmallisten liittymien ryhmittyminen lukumäärältään rajattuun ja kooltaan suhteellisen pieneen ryhmään on oleellista vianhallinnan kannalta, jotta menetelmän tulosten käyttö olisi taloudellisesti kannattavaa.

Seuraavaksi kuvataan lyhyesti lopputyössä käytetyt työkalut ja tarkempi aineisto. Tämän jälkeen tutkitaan kokeellisesti trajektorianalyysin tuloksia ensin vaihtelemalla mallin polynomiastetta kuuden kuukauden aikasarjalla ja seuraavaksi lyhennetyllä kolmen kuukauden aikasarjalla. Tämän jälkeen tutkitaan erikseen tiketillisiä ja tiketittömiä liittymiä ja etsitään tietoa liittymämäärän vaiku-

tuksesta trajektorianalyysin tuloksiin. Lopuksi ennustetaan uusien mittausten ryhmittymistä.

## 4.1 Työkaluvalinnat

Työssä käytetyn PC-laitteiston ja perusohjelmistojen tiedot ovat:

Lenovo ThinkPad W540
Windows 7 Enterprise
Service Pack 1
Intel® Core™ i7-4800MQ CPU @ 2.70 GHz
16,0 GB RAM
64 bit OS.

Tilastolliseen analysointiin käytetyt erikoisohjelmistot ovat:

64-bit R 3.1.1
flexmix 2.3-13 (Grün et al. 2015)
mgcv 1.8-6. (Wood et al. 2015)

## 4.2 Analysoitavan mallin ja aineiston rajaus

Koska *flexmix* ei kykene käsittelemään käytetyllä laiteympäristöllä koko suurta aineistoa yhtäaikaisesti, siitä valitaan aina jokin rajattu mielenkiintoinen osa-alue kerralla tarkasteltavaksi.

Trajektorianalyyseissä käytetään aina neljää vastemuuttujaa: vaimennukset ja kohinamarginaalit sekä paluu- että myötäsuuntiin. Analyyseissä käytetty trajektorimalli on täten aina samankaltainen polynomiasteen vaihdellessa tarpeen mukaan. Esimerkiksi kolmen polynomiasteen malli on

$$y_{it}^{*} = \beta_0^j + \beta_1^j Mittauspv_{it} + \beta_2^j Mittauspv_{it}^2 + \beta_3^j Mittauspv_{it}^3 + \varepsilon_{it}, \quad (4.1)$$

jossa normaalijakaumaa noudattavat vasteet $y_{it}^{*}$ ovat log ($Vaimennus\_p_{it}$), $Kohinamarginaali\_p_{it}$, log($Vaimennus\_m_{it}$) ja $Kohinamarginaali\_m_{it}$.

### 4.2.1 Aineisto A

Suurimpaan osaan lopputyön analyysejä on aineistoksi valittu maantieteellisesti rajattu osa-alue, joka sisältää yhteensä 305353 mittausta 1811 liittymästä kuuden kuukauden tarkastelujaksolla taulukossa 4.1. esitetyin tunnusluvuin. Analyyseissä tutkitaan muun muassa polynomiasteen, aikasarjan pituuden ja otoksen liittymämäärien vaikutusta trajektorianalyysin mallinnukseen. Tikettejä on yhteensä 55.

**Taulukko 4.1.** Mittausaineiston A muuttujat.

| Muuttuja | Mediaani/Keskiarvo | Min/Max |
|---|---|---|
| Nopeus_p | 1019/2111 | 256 / 11408 |
| Nopeus_m | 2464/6074 | 19/24574 |
| Vaimennus_p | 9.0/10.1 | 0.0/770.0 |
| Vaimennus_m | 16.1/17.9 | 0.0/375.0 |
| Kohinamarginaali_p | 8.8/11.3 | 0.0/320.0 |
| Kohinamarginaali_m | 19.3/21.4 | 0.0/425.0 |
| Tiketit/mittaus | 0.0/0.0002 | 0/2 |

### 4.2.2 Aineisto B

Osassa analyysejä käytetään tiketillisten liittymien aineistoa, mikä on saatu valitsemalla alkuperäisestä aineistosta (luku 3.1) kaikki ne liittymät, joilla on yksikin tiketti tarkastelujaksolla. Näitä liittymiä on yhteensä 732 kpl ja niillä on yhteensä 121158 mittausta taulukossa 4.2. esitetyin tunnusluvuin.

**Taulukko 4.2.** Mittausaineiston B muuttujat.

| Muuttuja | Mediaani/Keskiarvo | Min/Max |
|---|---|---|
| Nopeus_p | 2312/3693 | 256/11408 |
| Nopeus_m | 5704/8634 | 356/24575 |
| Vaimennus_p | 9.5/11.8 | 0.0/246.0 |
| Vaimennus_m | 14.0/16.9 | 0.0/495.0 |
| Kohinamarginaali_p | 11.4/12.2 | 0.0/300.0 |
| Kohinamarginaali_m | 11.3/14.2 | 0.0/385.0 |
| Tiketit/mittaus | 0/0.008 | 0/2 |

## 4.3 Polynomiasteen valinta

Seuraavaksi tutkitaan aineiston A avulla, miten polynomiasteen valinta vaikuttaa trajektorianalyysin tuloksiin ja mikä polynomiaste on paras joko muodostuvien ryhmien tai BIC:n perusteella.

### 4.3.1 Analyysiajojen numeerisia tuloksia

Analyysiaineistolle tehdään trajektorianalyysi toistaen vaihdellen polynomiastetta neljästä seitsemään. Analyysiajoista kerätään taulukossa 4.3. esitettyjä perustietoja: suppenemiseen vaadittu iteraatioiden lukumäärä, ajon kesto minuutteina, tuloksena syntyvien ryhmien lukumäärä ja ryhmien sisällä olevien liittymien

lukumäärä ja tiketit. Kerättyjen tietojen perusteella haetaan erityisesti lisäymmärrystä siitä, miten mallin polynomiasteen valinta vaikuttaa saatuihin tuloksiin.

**Taulukko 4.3.** Trajektorianalyysien suoritustietoja.

| Polynomiaste | Ryhmät: Liittymät | Tiketit ryhmittäin | Iteraatiot | Aika |
|---|---|---|---|---|
| 4 | 2: 450/1361 | 8/47 | 14 | 17 |
| 4 | 2: 450/1361 | 8/47 | 9 | 11 |
| 4 | 1: 1811 | 55 | 3 | 5 |
| 4 | 2: 450/1361 | 8/47 | 17 | 20 |
| 4 | 1: 1811 | 55 | 3 | 5 |
| 4 | 2: 450/1361 | 8/47 | 10 | 13 |
| 4 | 2: 450/1361 | 8/47 | 12 | 14 |
| 5 | 2: 450/1361 | 8/47 | 13 | 16 |
| 5 | 3: 93/864/854 | 0/34/21 | 22 | 26 |
| 5 | 1: 1811 | 55 | 3 | 5 |
| 5 | 1: 1811 | 55 | 3 | 5 |
| 5 | 3: 93/864/854 | 0/34/21 | 33 | 38 |
| 5 | 2: 450/1361 | 8/47 | 14 | 17 |
| 5 | 2: 450/1361 | 8/47 | 8 | 10 |
| 6 | 2: 450/1361 | 8/47 | 8 | 11 |
| 6 | 3: 91/919/801 | 0/27/28 | 14 | 17 |
| 6 | **3: 92/872/847** | **0/24/31** | 19 | 32 |
| 6 | 2: 450/1361 | 8/47 | 20 | 23 |
| 6 | 2: 450/1361 | 8/47 | 16 | 14 |
| 6 | 2: 450/1361 | 8/47 | 8 | 11 |
| 6 | 2: 450/1361 | 8/47 | 8 | 10 |
| 7 | 2: 450/1361 | 8/47 | 20 | 24 |
| 7 | 2: 450/1361 | 8/47 | 12 | 15 |
| 7 | 1: 1811 | 55 | 3 | 5 |
| 7 | 2: 450/1361 | 8/47 | 18 | 22 |
| 7 | 2: 450/1361 | 8/47 | 33 | 38 |
| 7 | 2: 707/1104 | **4/51** | 24 | 27 |
| 7 | 2: 450/1361 | 8/47 | 25 | 28 |

Vaikka toistoja on suhteellisen pieni määrä, voidaan tehdä viitteellisiä johtopäätöksiä käytetyn aineiston osalta. Kaksi ryhmää lähes aina samalla sisällöllä osoittautui yleisimmäksi trajektorianalyysin tulokseksi polynomiasteilla neljästä seitsemään. Kolmen ryhmän tulos oli mahdollinen viidennen ja kuudennen polynomiasteen analyyseissä. Yhden ryhmän tulos oli mahdollinen neljännen, viidennen ja seitsemännen polynomiasteen analyyseissä. Tästä voinee vetää yleisemmän johtopäätöksen, että polynomiasteen kasvaessa ryhmien lukumäärän kasvu on todennäköisempää, mutta kun polynomiaste kasvaa entisestään, ryhmien määrä alkaa keskimäärin vähenemään. Käytetyllä aineistolla kolme ryhmää osoittautui suurimmaksi mahdolliseksi tulokseksi, vaikka R-koodissa tavoiteltiin viittä.

Analyysiajon kestoon vaikuttaa pääosin ennen suppenemista tarvittavien iteraatioiden määrä. Ryhmien lukumäärän kasvaessa tarvittavien iteraatioiden määrä tyypillisesti kasvaa. Toisin sanoen, mitä useampi ryhmä, sitä pitempi analyysin suoritusaika on odotettavissa.

Samalla ryhmälukumäärällä saadaan tyypillisesti joko täysin samanlainen lopputulos (yksi ja kaksi ryhmää sekä viidennen polynomiasteen tuottama kolme ryhmää) tai lähes sama lopputulos (kuudennen polynomiasteen kolme ryhmää). Kolmen ryhmän välillä tapahtuu täten pientä siirtymää yhtälön (2.2) mukaisten pienempien posteriori-todennäköisyyksien välillä sen mukaan, mitä satunnaislukuja valikoituu prosessiin. Tarkasteluaineistolla viidennen polynomiasteen malli keskitti hieman paremmin tikettejä samaan ryhmään verrattuna kuudennen polynomiasteen tulokseen. Täten vianhallinnan näkökulmasta viisi lienee hieman parempi kuin kuusi polynomiastetta, mutta saaduilla tuloksilla ero ei liene kuitenkaan merkityksellinen. Sen sijaan eräs kahden ryhmän analyysiajo seitsemännen polynomiasteen mallilla erottelee poikkeuksellisen hyvin tiketilliset liittymät (4/51) suurempaan ryhmään. Jaottelu olisi vielä hyödyllisempi, jos ryhmien kokosuhde olisi esimerkiksi päinvastoin, nyt suurempaan ryhmään kuuluu noin 61% kaikista liittymistä.

Taulukossa 4.4. esitettyjen BIC-arvojen perusteella kuudennen polynomiasteen malli kolmella ryhmällä osoittautui parhaaksi, kun uusia analyysiajoja tehtiin niin kauan, kunnes kaikki yhdistelmät esiintyivät. Kolmen ryhmän esiintyminen neljännen ja seitsemännen polynomiasteen mallissa osoittautui haasteelliseksi, mutta on mahdollinen riittävän monen analyysiajon avulla. Seitsemännen polynomiasteen trajektorianalyysi, jossa tiketit erottuvat hyvin kahden ryhmän välillä, osoittautui heikoksi BIC-näkökulmasta.

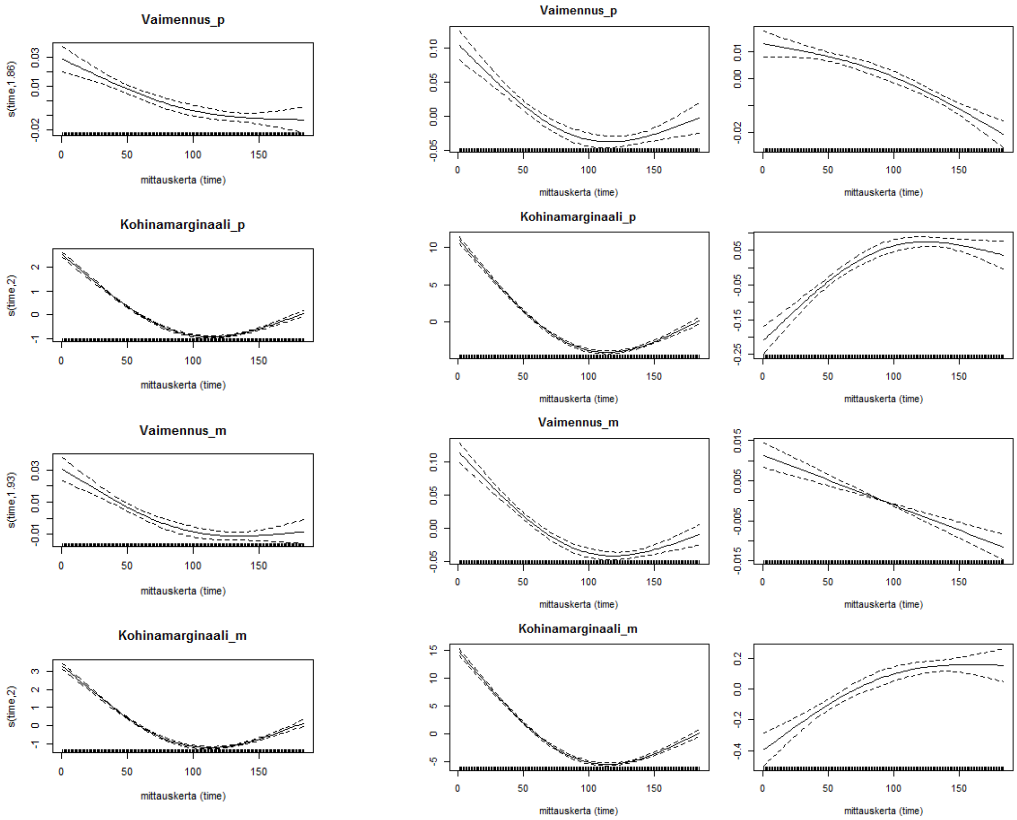
**Taulukko 4.4.** Trajektorianalyysien BIC-arvoja.

| Polynomiaste | 1 ryhmä | 2 ryhmää | 3 ryhmää |
|---|---|---|---|
| 4 | 6608081 | 5376377 | 4807039 |
| 5 | 6607970 | 5376259 | 4806461 (0/24/31)<br>4806663 (0/21/34) |
| 6 | 6607792 | 5376066 | **4805958 (24/0/31)**<br>4806208 (0/27/28)<br>4806377 (0/21/34) |
| 7 | 6607774 | 5762770 (4/51)<br>5376076 (8/47) | 4805971<br>4806208 (0/27/28) |

### 4.3.2 Trajektorit eri ryhmämäärillä

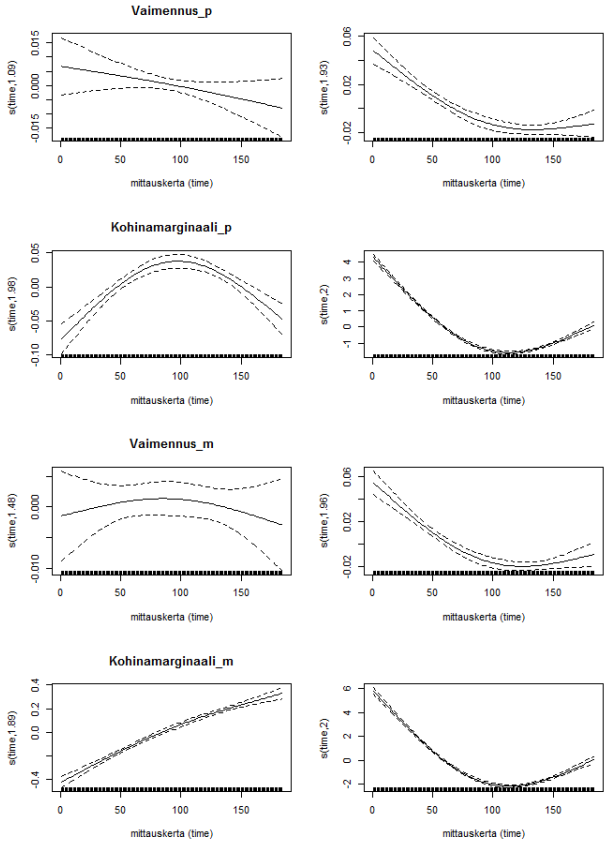
Seuraavaksi (kuviot 4.1.- 4.3.) esitellään eri ryhmälukumäärien tuottamat monitrajektorit, jossa vaaka-akselin numerot kertovat tarkastelupäivien järjestysluvut. Yhden ryhmän trajektorit ovat muodoltaan samantapaiset kuin kahden ryhmän vasemman puoleinen pienempi ryhmä, mutta käyrän muodon voimakkuus on selvästi laimeampi. Yhden ryhmän trajektorit eivät lisää aineiston ymmärrystä,

koska saatu tulos on yksinkertaisesti liian lattea. Kahden ryhmän tapauksessa pienemmällä ryhmällä on vähemmän tikettejä suhteessa ryhmässä oleviin liittymämääriin (1.8%) kuin isommalla ryhmällä (3.5%). Näistä isompi ryhmä on toimialatietämyksen perusteella se odotetumpi ja toivottukin käyttäytyminen: vaimennusten laskiessa kohinamarginaalit kasvavat. Pienemmän ryhmän kaikilla muuttujilla on laskeva trendi ja lopussa hieman nousua. Tämänkaltainen käyttäytyminen on jo hieman yllätyksellisempää reaalimaailman näkökulmasta, mutta kyse lienee tyypillisesti hyvin pienistä nopeuksien laskuista, jolloin tulee tilaa sekä vaimennuksen että kohinamarginaalien laskuille kokonaisuuden toimiessa samalla luotettavammin.



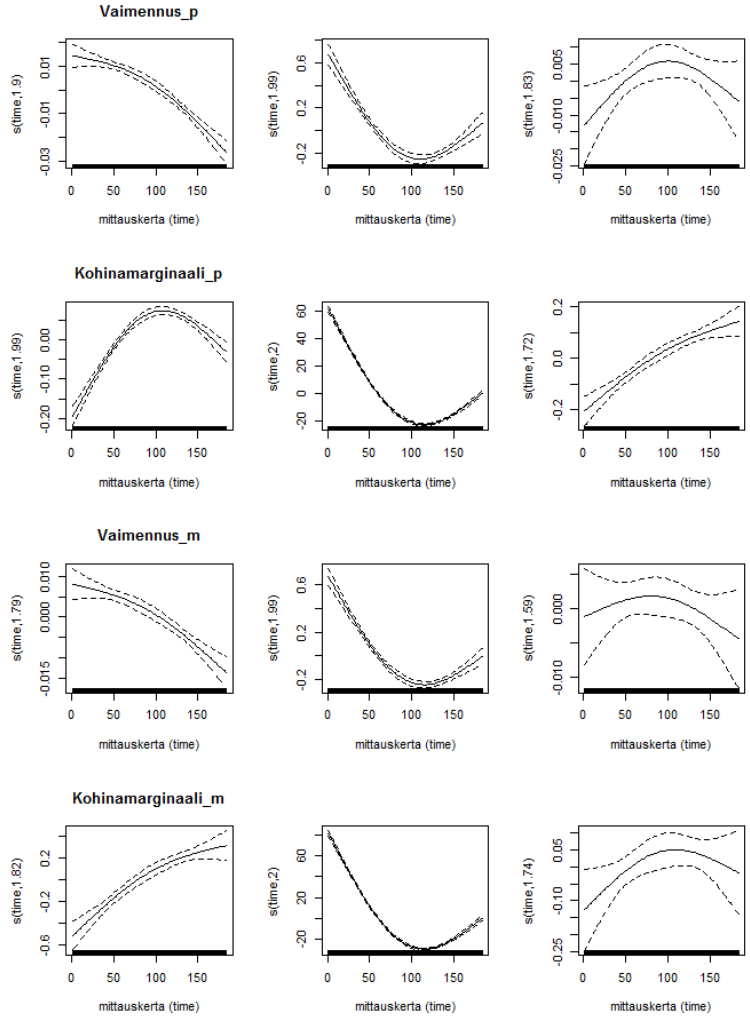

**Kuvio 4.1.** Yhden ja kahden ryhmän (450/1361 liittymää) monitrajektorit.

Lukuisissa seitsemännen polynomiasteen analyysiajoissa ilmaantui kertaalleen kuviossa 4.2. esitetty ryhmittely, jossa tiketit (4/51) keskittyvät pääosin oikeanpuoleiseen, suurempaan ryhmään. Ryhmien vikaantuminen suhteessa liittymien lukumäärään on 0.57% ja 4.6%.




**Kuvio 4.2.** Kahden ryhmän (707/1104 liittymää) monitrajektorit.

Kolmen ryhmän tuloksissa liittymien lukumäärissä oli enemmän vaihtelevuutta, mutta käytännössä trajektorit ovat esimerkin kaltaiset ja riittää, että tarkastellaan vain yhtä tapauksista. Kolmen ryhmän kuviossa 4.3. pienin keskimmäinen ryhmä vastaa muodoltaan yhden ja kahden ryhmän trajektoria, käyrän voimakkuus on vain suurempi ja ryhmän koko huomattavasti paljon pienempi (noin 5% liittymistä). Samalla pienin ryhmä on erittäin hyvin toimiva; siihen kuuluvilla liittymillä ei ole yhtään tikettiä tarkastelujaksolla.

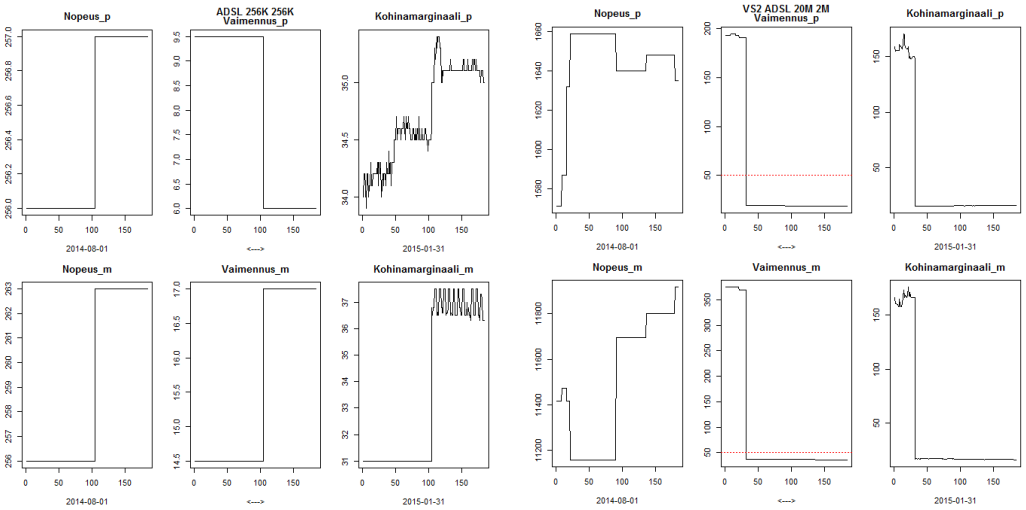



**Kuvio 4.3.** Kuudennen polynomiasteen (872/92/847 liittymää) monitrajektorit.

Oikeanpuoleisen ryhmän paluusuunnan kohinamarginaali on nouseva ja paluusuunnan vaimennus parabolinen, muut trajektorit ovat vakioita 95% luotettavuusvälillä. Muissa kuin kuudennen polynomiasteen (parhaiden BIC-arvojen) malleissa paluusuunnan vaimennus on vakio. Ryhmän tikettien määrä suhteessa siinä olevien liittymien lukumäärään (3.7%) on suurin kaikista ryhmistä. Tämä trajektorimuotojen yhdistelmä tulee esiin vain kolmen ryhmän tuloksessa. Vasemmalla sijaitsevan ryhmän vikaantuvuus on 2.8% suhteessa siihen kuuluvien liittymien lukumäärään. Trajektorimuoto vastaa kahden ryhmän mallin suurempaa ryhmää, mikä on samalla oletuskäyttäytyminen. Kooltaan vasemmalta katsoen ensimmäinen ja kolmas ryhmä ovat noin saman suuruiset.

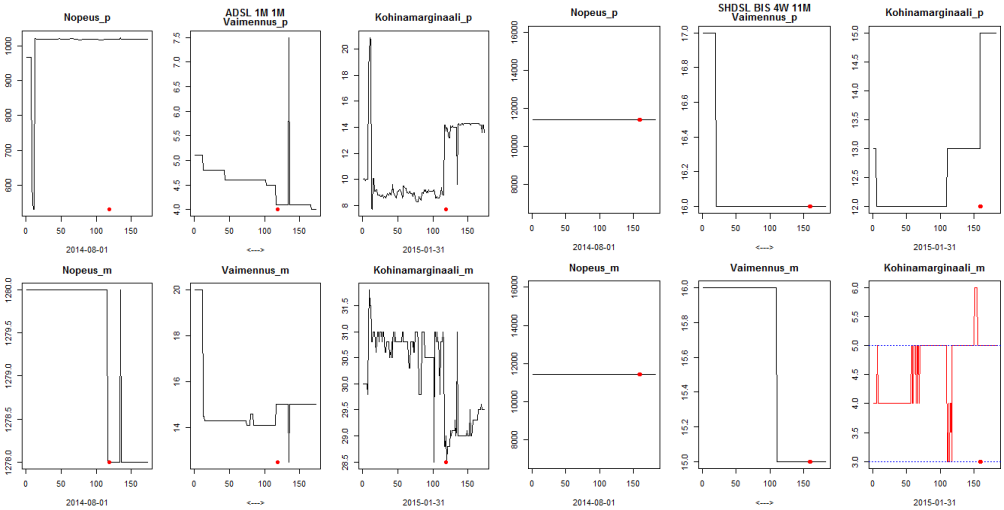
Kuvioissa 4.4. - 4.6. on esitetty kuvion 4.3. klustereihin kuuluvien liittymien aikasarjoja syvällisemmän ymmärryksen saamiseksi. Ensin on kuvattu kaksi keskimmäiseen ryhmään kuuluvaa liittymää, joilla ei ole yhtään tikettiä. Punainen väri kuvaa mahdollisesti ongelmallista muuttujan arvon tasoa, mutta

tähän tulee suhtautua erittäin kriittisesti. Nimittäin aikasarjoja laajemmin tutkittaessa selvisi, että yksittäisen muuttujan absoluuttista kriittistä tasoa on erittäin vaikea uskottavasti määritellä, koska arvon hyväksyttävyys on suuresti riippuvainen liittymän kokonaistilanteesta.



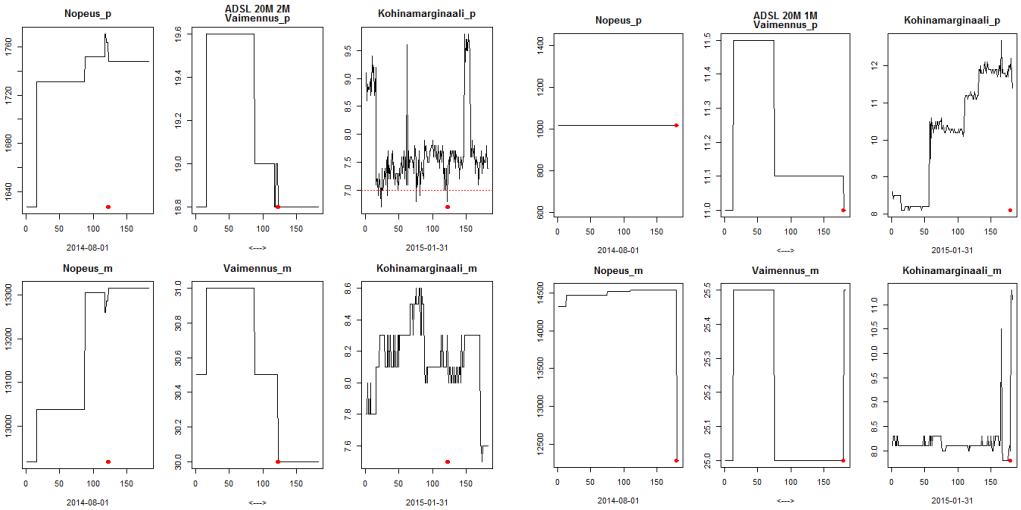

**Kuvio 4.4.** Keskimmäiseen ryhmään kuuluvien liittymien aikasarjoja.

Seuraavaksi kuviossa 4.5. esitellään kaksi vasemmanpuoleiseen ryhmään sisältyvää liittymää. Näiden aikasarjojen ja olemassa olevan tikettitiedon perusteella vaikuttaa siltä, että muuttujissa tapahtuvat tasomuutokset aiheuttavat osassa laajakaistaliittymiä tilapäistä häiriötä laajakaista-modeemin etsiessä stabiilia tilaa muuttuneessa tilanteessa. Dynaamiset syy- ja seuraussuhteet ovat moninaiset, mikä aiheuttaa suurta haastetta laadukkaalle ennakoivalle viankorjaukselle. Punainen merkkipallo graafissa ilmaisee, että kyseisenä mittauspäivänä käyttäjä on avannut yhden tai useamman tiketin kyseessä olevalle liittymälle.




**Kuvio 4.5.** Vasemmanpuoleisen ryhmän liittymien aikasarjoja.

Kuviossa 4.6. esitetään kaksi oikeanpuoleiseen ryhmään kuuluvaa liittymää, joilla on vähintään yksi tiketti tarkastelujaksolla. Tiketin alkusyynä näyttäisi olevan laitteistossa tapahtuva muutos, jonka eskaloima stabiilin tilan etsintä aiheuttaa tilapäisiä yhteysongelmia.




**Kuvio 4.6.** Oikeanpuoleisen ryhmän liittymien aikasarjoja.

Niin kuin edellisistä aikasarjagraafeistakin näkee, ryhmiin kuuluu hyvin monenlaisia liittymiä, jotka on valittu suurimman posteriori-todennäköisyyden perusteella. Vianhallinnan näkökulmasta on ongelmallista, että tiketin omaavia liittymiä ei saada paremmin kohdistettua parhaassa tapauksessa yhteen kooltaan mahdollisimman pieneen ryhmään, vaan tiketit jakautuvat valitettavan tasaisesti esimerkiksi kahden ryhmän tapauksessa suurempaan tai kolmen ryhmän tapauksessa kahteen suurimpaan pienimmän ryhmän toimiessa virheittä. Kun toimitaan satunnaisuuden ehdoilla, ryhmien lukumäärä vaihtelee ja tarvittaessa joudutaan ajamaan trajektorianalyysia kauan, kunnes on saatu haluttu määrä ryhmiä. Tämän jälkeen tulosten tarkastelu on jo kuitenkin ennakoidumpaa ainakin työssä tarkastellulla aineistoilla.

## 4.4 Mittauspisteiden lukumäärän vaikutus malliin

Kolmen kuukauden aineistolle (125840 mittausta 1478 liittymästä) tehdään trajektorianalyysi toistaen vaihdellen polynomiastetta neljästä seitsemään. Aineisto on saatu poistamalla aiemmin käytetystä kuuden kuukauden aineistosta A kolmen viimeisen kuukauden mittaukset. Saatuja tuloksia verrataan edellisessä luvussa 4.3 saatuihin tuloksiin.

### 4.4.1 Analyysiajojen numeerisia tuloksia

Analyysiajoista kerätään taulukossa 4.5. esitettyjä perustietoja: suppenemiseen vaadittu iteraatioiden lukumäärä, ajon kesto minuutteina, tuloksena syntyvien ryhmien lukumäärä ja ryhmän liittymien lukumäärä ja tiketit. Edellisten lisäksi kerätään talteen eri mallien BIC-tieto (taulukko 4.6.), jonka perusteella voidaan päätellä parhaiten aineistoon istuva malli. Koska kolmen kuukauden aineistossa on vain vähän tikettejä (tiketillisten liittymien osuus kaikista liittymistä on noin 0.7% ), vikaantumisen jakautumisen tarkastelu ei ole niin mielenkiintoista.

**Taulukko 4.5.** Trajektorianalyysien suoritustietoja.

| Polynomiaste | Ryhmät: Liit-tymät | Tiketit ryhmit-täin | Iteraatiot | Aika |
|---|---|---|---|---|
| 4 | 2:1093/385 | 7/3 | 10 | 4 |
| 4 | 1: 1478 | 10 | 3 | 2 |
| 4 | 2: 1093/385 | 3/7 | 13 | 6 |
| 4 | 2: 1103/375 | 7/3 | 21 | 18 |
| 4 | 1: 1478 | 10 | 3 | 2 |
| 4 | 2: 1103/375 | 7/3 | 10 | 4 |
| 4 | 1: 1478 | 10 | 11 | 4 |
| 5 | 1: 1478 | 10 | 3 | 2 |
| 5 | 2: 385/1093 | 3/7 | 12 | 4 |
| 5 | 2: 375/1103 | 3/7 | 19 | 7 |
| 5 | 2: 381/1097 | 3/7 | 8 | 3 |
| 5 | 2: 385/1093 | 3/7 | 7 | 4 |
| 5 | 2: 385/1093 | 3/7 | 15 | 6 |
| 5 | 3: 77/792/609 | 0/5/5 | 19 | 8 |
| 6 | 2: 375/1103 | 3/7 | 22 | 9 |
| 6 | 2: 385/1093 | 3/7 | 11 | 4 |
| 6 | 3:77/608/793 | 0/5/5 | 18 | 7 |
| 6 | 3: 77/608/793 | 0/5/5 | 16 | 7 |
| 6 | 1: 1478 | 10 | 9 | 4 |
| 6 | 1: 1478 | 10 | 3 | 2 |
| 6 | 3: 77/608/793 | 0/5/5 | 22 | 9 |
| 7 | 1: 1478 | 10 | 3 | 2 |
| 7 | 3: 77/608/793 | 0/5/5 | 22 | 9 |
| 7 | 2: 385/1093 | 3/7 | 25 | 10 |
| 7 | 2: 385/1093 | 3/7 | 13 | 5 |
| 7 | 1: 1478 | 10 | 3 | 2 |
| 7 | 2: 385/1093 | 3/7 | 20 | 8 |
| 7 | 3: 77/608/793 | 0/5/5 | 19 | 9 |

Lyhemmällä mittausjaksolla analyysiajojen kesto verrattuna tarvittuihin iteraatioihin on selvästi alempi verrattuna kuuden kuukauden aineistoon. Tulosten vaihtoehdot ovat hieman suppeammat ja ryhmälukumäärien vaihtoehdot ovat

tasaisemmin edustettuina kuin pidemmällä aikasarjalla, jolla kahden ryhmän tulos oli yleisin.

**Taulukko 4.6.** Trajektorianalyysien BIC-arvoja.

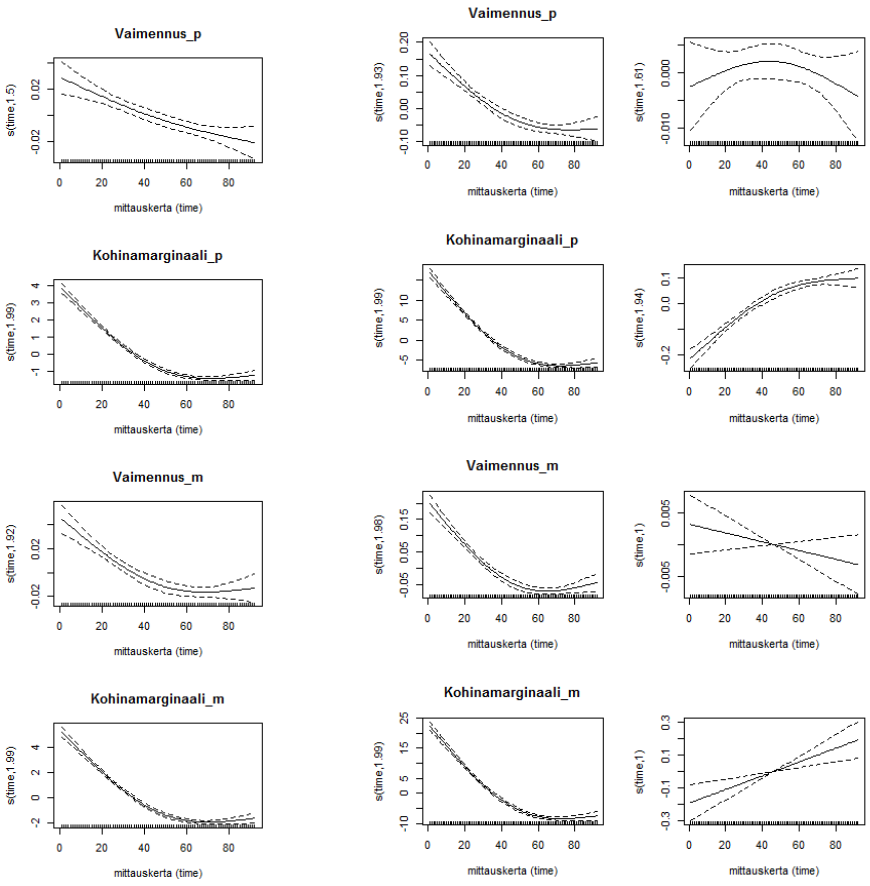
| Polynomiaste | 1 ryhmä | 2 ryhmää | 3 ryhmää |
|---|---|---|---|
| 4 | 2928337 | 2190807 | - |
| 5 | 2928380 | 2190891<br>2190908 | **1919256** |
| 6 | 2928357 | 2190898<br>2190902 | 1919261 |
| 7 | 2928361 | 2190943 | 1919320 |

BIC-arvoista puuttuu kokonaan neljännen polynomiasteen kolmen ryhmän mallin tieto, koska sellaista tulosta ei ilmaantunut lukuisten toistojenkaan jälkeen. Näin ollen BIC:n perusteella parhaaksi aineistoa sovittavaksi malliksi tuli viidennen polynomiasteen kolmen ryhmän malli. Täten yhtä polynomiastetta pienempi malli yltää parhaaseen tulokseen aiempaa puolta pienemmällä mittausjaksolla.

### 4.4.2 Trajektorit eri ryhmämäärillä

Seuraavaksi (kuviot 4.7. – 4.8. ) esitellään eri ryhmämäärien tuottamat monitrajektorit. Ensin yhden ja kahden ryhmän mallit, missä toinen multitrajektori on saman muotoinen kuin kuuden kuukauden aineistolla. Kahden ryhmän mallissa oikeanpuoleisen ryhmän trajektoriyhdistelmä, missä kohinamarginaalit ovat nousevia ja vaimennukset ovat vakioiset, on uudenlainen.

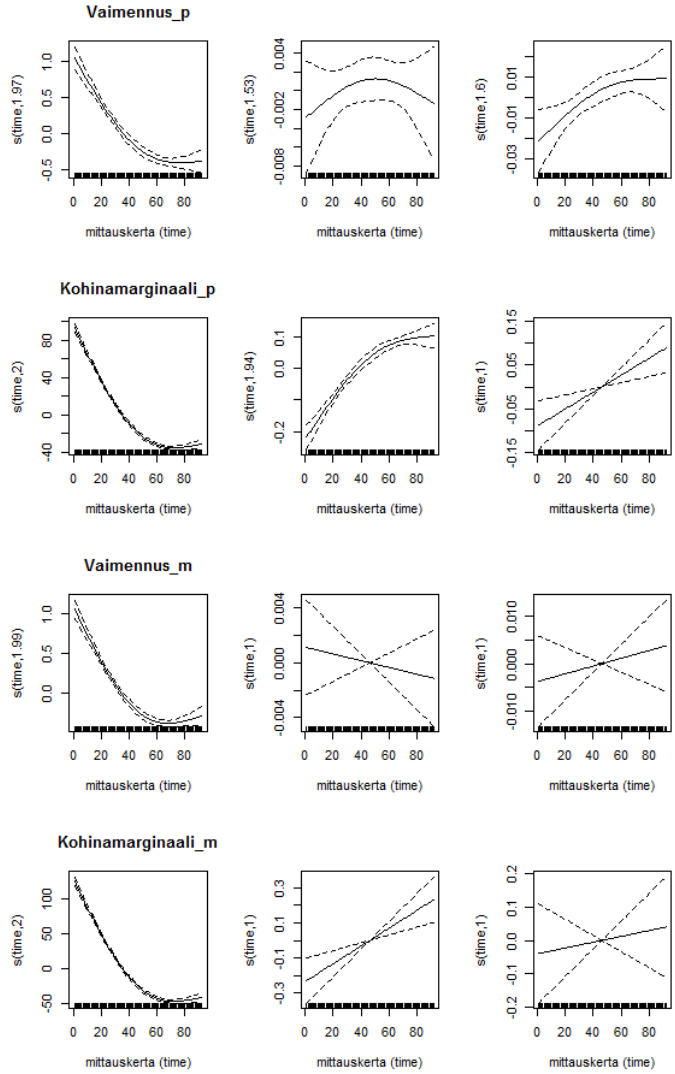
Kolmen ryhmän malli sisältää samat kahden ryhmän mallin muotoiset multitrajektorit ja niiden lisäksi uudenlaisen trajektoriyhdistelmän, missä paluusuunnan vaimennus ja kohinamarginaali ovat nousevia ja myötäsuunnan vakioiset 95% luottamusvälillä.




**Kuvio 4.7.** Yhden ja kahden ryhmän (450/1361) monitrajektorit.

Kun vertailee kuviossa 4.8. esitettyä tulosta ja kuuden kuukauden vastaavia kolmen ryhmän monitrajektoreita (kuvio 4.3.) pareittain (keskimmäinen verrattuna vasemmanpuoleiseen ja oikeanpuoleiset keskenään), niin vaikuttaa siltä, että pitemmän mittausjakson tuottama informaatio antaa enemmän ymmärrystä liittymän käyttäytymisestä kuin osittain torsoksi osoittautunut kolmen kuukauden esitys. Yksi kolmesta ryhmästä on muodoltaan noin sama molemmissa malleissa; kaikki trajektorit ovat parabolisesti laskevia ja lopussa lievästi nousevia. Lopuissa kahdessa ryhmässä huomaa, että kun tarkastelujakso on vain puolet maksimista, niin muuttujan käyttäytymisestä ei voi tehdä niin luotettavia päätelmiä kuin kuuden kuukauden aineistolla. Esimerkiksi vaimennuksen paluusuunta on lyhemmällä aikajaksolla vakioinen, kun kuuden kuukauden aikajaksolla trajektori on lineaarisen laskeva. Toisella ryhmäparilla taasen paluusuunnan vaimennus näyttäisi olevan lineaarisen nouseva, mutta pitemmällä aikajaksolla se onkin ylöspäin aukeava paraboli. Pidemmällä aikasarjalla saadaan täten samalla tietoa siitä, missä ajankohdassa suunnanvaihdos tapahtuu ja onko se merkitsevä 95% luottamusvälillä. Toisaalta ei näyttäisi olevan mitään syytä, miksi tarkastelujaksoa pitäisi tästä enää pidentää, sillä se tarkoittaisi samalla pi-

dentyviä analyysiajoja. Liite D sisältää ryhmiin kuuluvien liittymien aikasarja-graafeja.



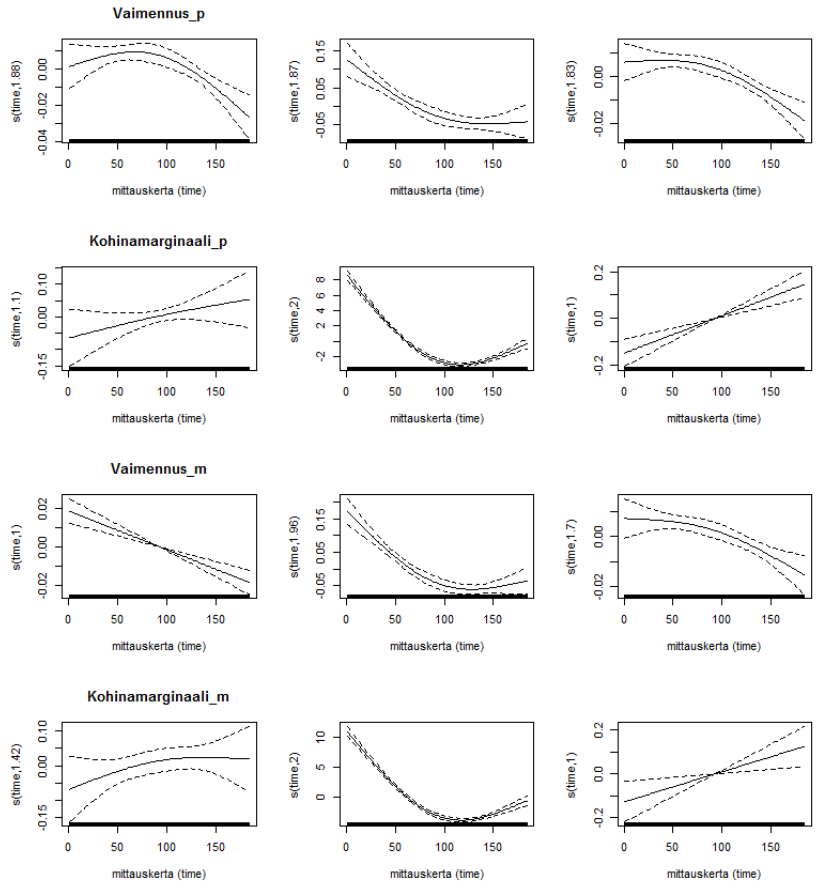

**Kuvio 4.8.** Kolmen ryhmän (77/792/609) monitrajektorit.

## 4.5 Tiketillisten liittymien trajektorimalleja

Seuraavaksi tutkitaan tarkemmin liittymiä, joilla on vähintään yksi tiketti tarkastelujaksolla (aineisto B). Enää ei olla niinkään kiinnostuneita mallinvalintaan johtavasta prosessista vaan pikemminkin lopputuloksesta ja sen antamasta tikettien syntymiseen liittyvästä lisätietämyksestä. Yleisenä huomiona mainittakoon, että kun liittymämäärä on noin 40% (732) edellä tutkitun aineisto A:n liittymämäärästä (1811), niin trajektorimallien tuottamat ryhmälukumäärät ja erilai-

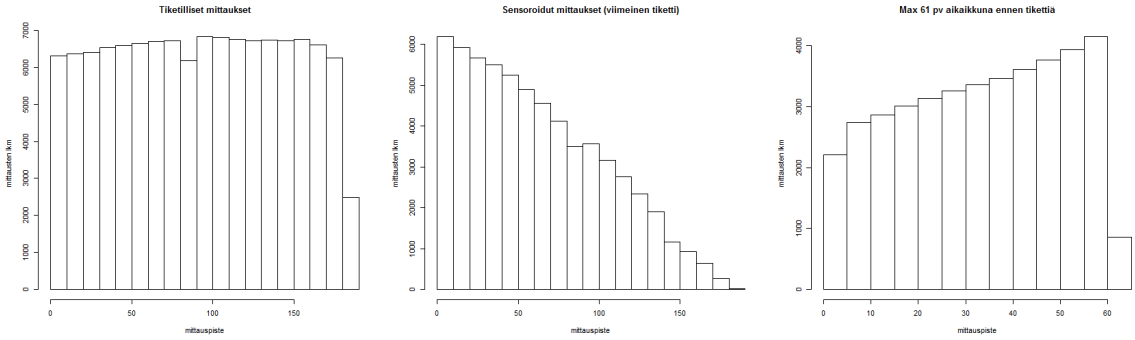
set sisällölliset vaihtoehdot ovat selvästi monipuolisempia ryhmämäärien vaihdellessa tyypillisesti kolmesta seitsemään.

Esitetyn mallin valintaan vaikuttivat sekä subjektiivinen trajektoritulosten tarkastelu että mallin tuottama BIC-arvo saman polynomiasteen erilaisiin tuloksiin verrattuna. Pienemmillä ryhmämäärillä trajektoreiden muodot pelkistyvät hyvin yleisiksi, eivätkä siten anna tarpeellisella tasolla uutta tietämystä ollakseen hyödylliset. Seuraavassa 4.9. kuviossa on tästä esimerkki. Toisaalta, kun ryhmälukumäärät kasvavat viimeistään viidestä ylöspäin, niin tyypillisesti ryhmien välisessä käyttäytymisessä mennään jo varsin hienovaraisiin eroihin, mikä ei enää yleisesti ottaen ole kovin hyödyllistä. Täten vaikka BIC:n perusteella aineisto istuisi paremmin näihin runsaslukuisempiin ryhmämalleihin, niin toimialatietämyksen perusteella löytynee kuitenkin parhaiten kuhunkin tilanteeseen hyödyllisin ryhmälukumäärä, jonka tuottama tietämys on riittävällä tasolla ja jonka ihminen pystyy vielä hyvin käsittelemään. Toisaalta, jos automatiikka käyttää saatua mallia, niin ryhmien lukumäärä sinänsä ei ole niin ratkaiseva, kunhan sisältö vain tukee sille asetettuja tavoitteita riittävällä tasolla. Kun tutkii 4.9. kuvion muodoltaan pelkistettyjä ryhmiä, niin ne muistuttavat muotokieleltään aiempia aineistolle A tehtyjä kolmen ryhmän trajektoreita.




**Kuvio 4.9.** Tiketillisten liittymien pelkistettyjä ryhmiä.

Seuraavaksi trajektorianalyysin tuloksia esitellään aineiston B avulla siten, että ensin tarkastellaan koko aineistoa B (121158 mittausta), jonka jälkeen muunnetaan se sensoroiduksi aineistoksi (62345 mittausta) poistamalla kaikki viimeisen tiketin jälkeiset mittaukset. Tämän jälkeen tehdään sensoroidusta aineistosta vielä katkaistu aineisto (40384 mittausta), jossa jokaista tikettiä mahdollisesti edeltävät 61 mittausta otetaan mukaan tarkasteluun ja aikaikkunan pituus toimii yhteisenä trajektoreiden x-akselina. Seuraavat histogrammit kuvaavat näiden erilaisten aineistojen mittausmääriä.

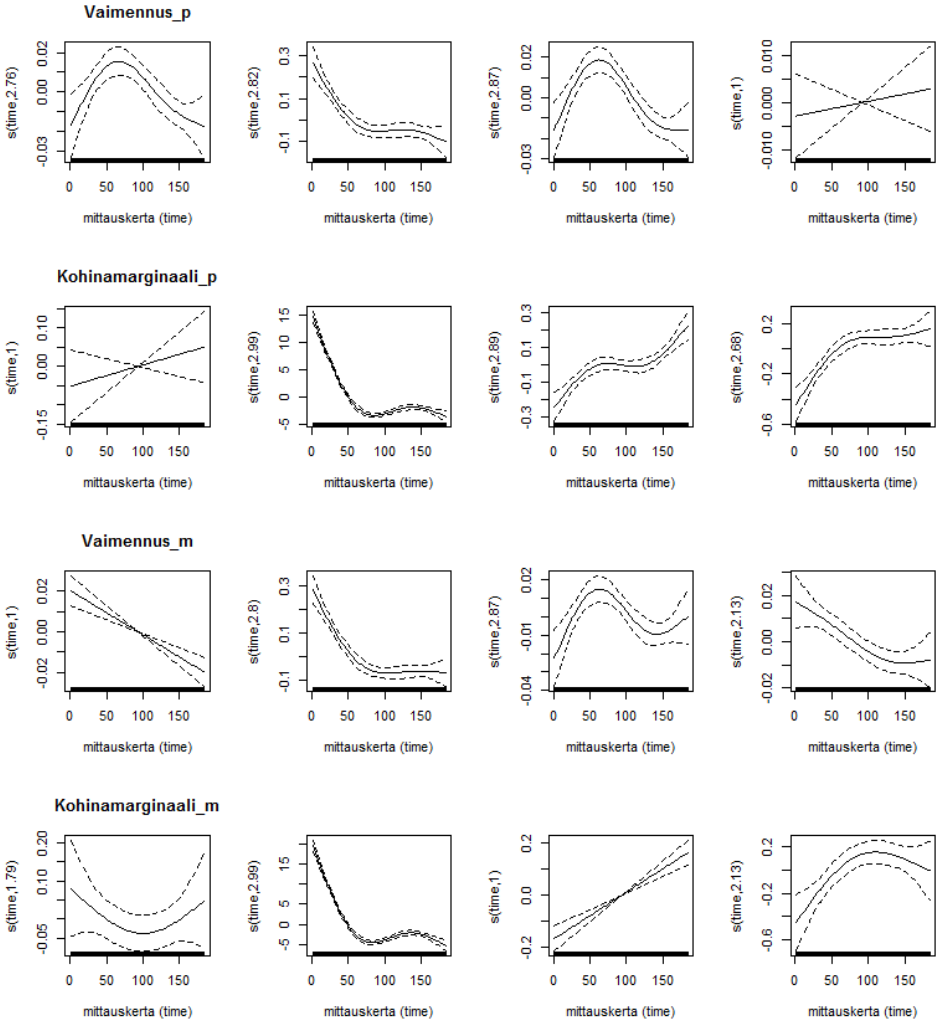



**Kuvio 4.10.** Tiketillisten liittymien koko/sensoroitu/katkaistu aineisto.

### 4.5.1 Tiketilliset liittymät koko tarkastelujaksolla

Kuuden kuukauden mittausaineistolla informatiivisimmaksi trajektorimalliksi osoittautui neljännen polynomiasteen neljän ryhmän malli, jonka trajektorit on esitetty kuviossa 4.11. Vasemmalta oikealle ryhmien koot ovat 197, 98, 242 ja 195 ja tikettien lukumäärät ovat 251, 119, 298 ja 245. Selvästi muita pienempi ryhmä on täten toinen vasemmalta, jossa kaikki muuttujat ovat laskevia. Saman ryhmän liittymillä on myös suhteessa hieman vähemmän tikettejä, mutta erot ovat varsin pienet. Vasemmanpuoleisen ryhmän liittymien vaimennukset laskevat kohti tarkastelujakson loppua muiden muuttujien ollessa vakioita 95% luottamusvälillä. Kahden oikean puoleisen ryhmän kohinamarginaalit ovat nousevia. Suurin ero näiden kahden ryhmän välillä on vaimennuksessa, joka on vakio paluusuuntaan ja laskeva myötäsuuntaan oikeanpuoleisessa ryhmässä, viereisessä ryhmässä paluusuunnan vaimennus on parabolisen laskeva ja myötäsuunnan ensin parabolisen laskeva ja sen jälkeen uudelle tasolle vakioituva.

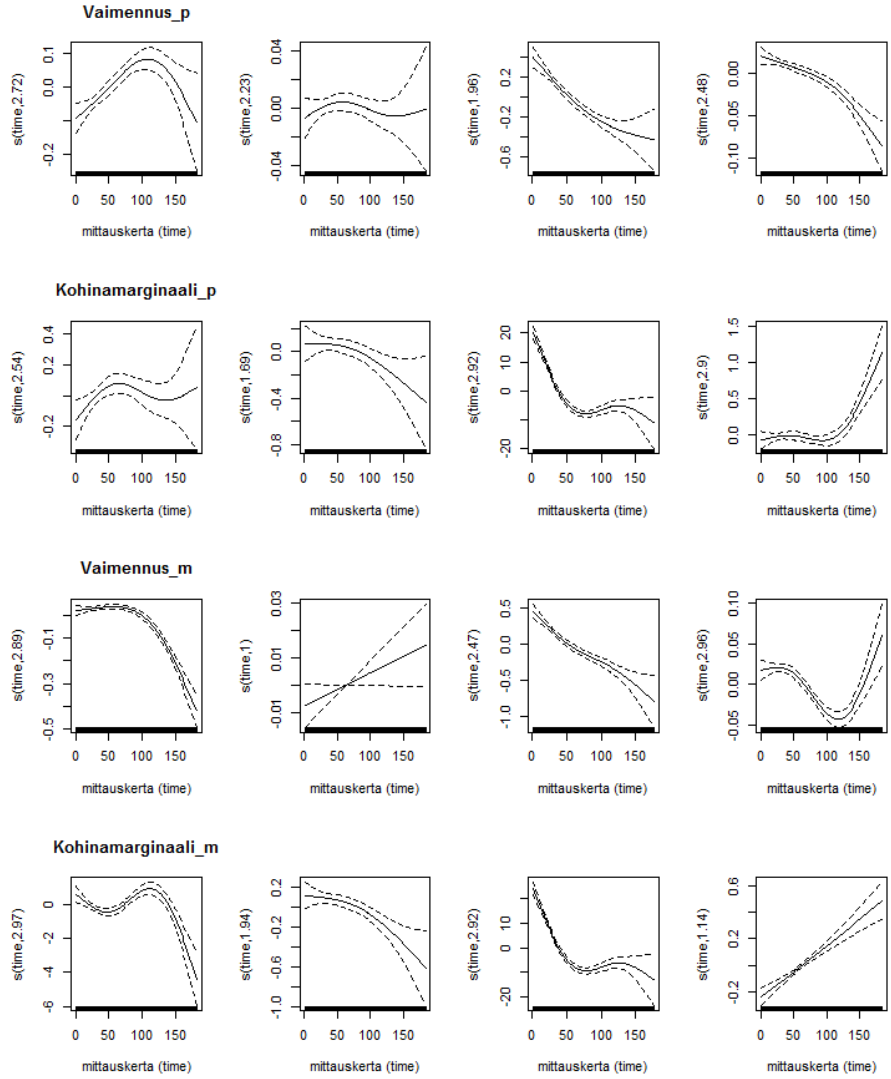
Näistäkin klustereista voi havaita, että tiketillisten liittymien klustereissa on tarjolla monta käyttäytymisen kehityksellistä vaihtoehtoa. Toisaalta useimmilla liittymillä on vain yksi tiketti pitkähköllä kuuden kuukauden tarkastelujaksolla, joten vikaantumiseen johtavasta kehityspolusta ei voine vielä tehdä yleisempää päätelmää.




**Kuvio 4.11.** Tikettejä omaavien liittymien trajektoreita.

### 4.5.2 Sensoroidut tiketilliset liittymät

Seuraavaksi aineistoa B muutetaan siten, että liittymän viimeisen tiketin jälkeiset mittaukset poistetaan kokonaan tarkastelusta, jolloin mittausten kokonaislukumäärä kohti tarkastelujakson loppua on laskeva. Sensoroidulla aineistolla on mahdollista tarkastella kehitystä ennen mielenkiintoista tapahtumaa ottamatta kantaa siihen, mitä sen jälkeen tapahtuu.




**Kuvio 4.12.** Sensoroitujen tikettiliittymien trajektoreita.

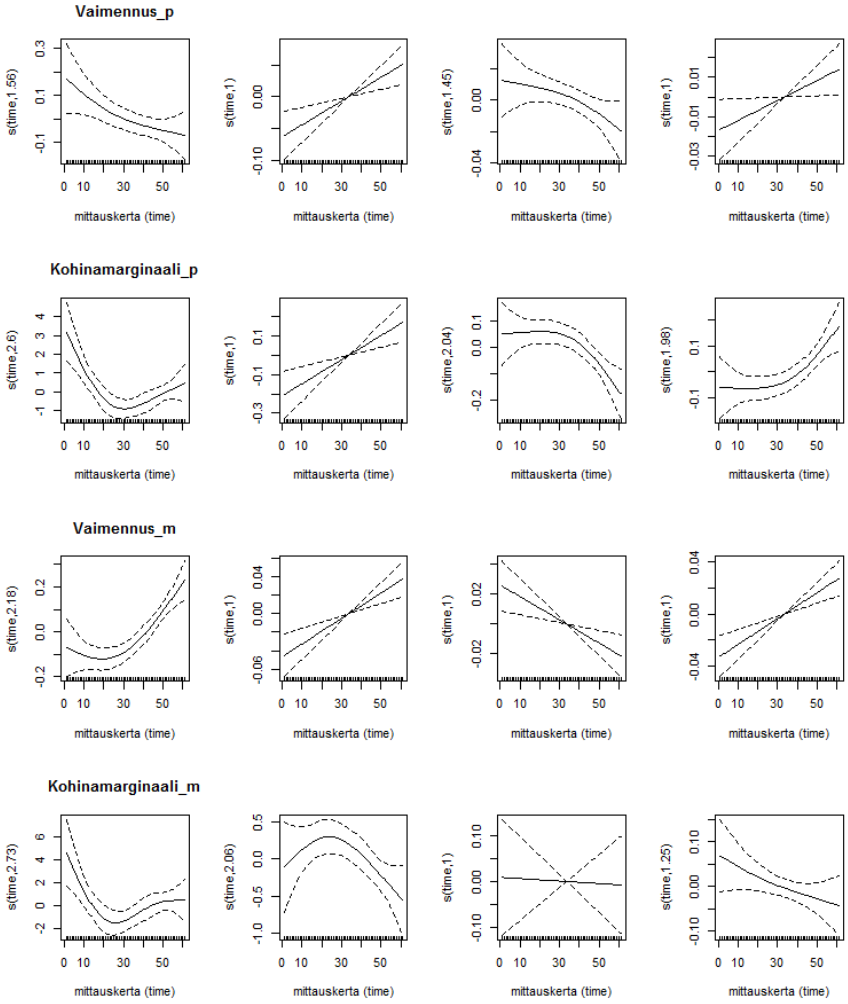
Parhaaksi malliksi valikoitui edellisessä kuviossa esitetty neljännen polynomiasteen ja neljän ryhmän malli, jonka ryhmien koot ovat 126, 210, 66 ja 330 ja tiketit ovat 154, 263, 82 ja 414. Tikettien suhteelliset osuudet ovat noin samat eri klustereissa, mutta oikeanpuoleisin ryhmä on selvästi suurin ja sen vieressä oleva on selvästi pienin kooltaan. Toisin sanoen, oikeanpuoleisimman ryhmän kehityskulku näyttäisi olevan tyypillisin ennen vikaantumista eli tilanne, jossa paluusuunnan vaimennus on laskeva ja muut muuttujat kasvavia. Pienimmän ryhmän kohdalla kaikki muuttujat ovat laskevia ja loppuvaiheessa asettuvat uudelle tasolle. Tässä kohtaa on kuitenkin hyvä huomata se seikka, että osalla liittymistä esiintyy useampiakin tikettejä, mikä voi osaltaan hieman vääristää tuloksia.

### 4.5.3 Katkaistut tiketilliset liittymät

Katkaistulla aineistolla poistetaan kunkin tiketin jälkeiset mittaukset ja tarkasteluun otetaan mukaan vain valitun aikaikkunan sisällä olevat mittaukset ennen tikettiä. Jos liittymällä on useampia eri päiviin sijoittuvia tikettejä, kukin tikettipäivä otetaan aina kerrallaan aikaikkunan loppupisteeksi. Liittymien lukumäärä on lopulta 854, mikä on suurempi kuin reaalimaailman tiketillisten liittymien lukumäärä (732). Tämä johtuu siitä, että jokaisella aineistossa olevalla aikasarjan päättävällä tiketillä tulee olla oma yksilöity liittymätunniste, jolle katkaistu aikasarja rakennetaan ennen analyysiä. Valitun aikaikkunan sisällä olevat tiketilliset mittaukset käsitellään samaan tapaan kuin mikä tahansa mittaus. Tästä syystä aineiston kokonaistikettimäärä kasvaa hieman, kun aikaikkunat menevät osittain päällekkäin saman liittymän eri tikettejä tarkasteltaessa erikseen. Tarkastelujaksoksi tulee valittu aikaikkuna, jolloin tikettien esiintyminen osuu aina aikaikkunan viimeiselle arvolle ja sitä ennen olevat mittaukset edustavat historiaa ennen tikettiä.

Trajektorianalyysiajojen tulosten perusteella näyttäisi siltä, että tällä aineistolla ja neljän muuttujan mallilla aikaikkunan pituus tulee olla vähintään kaksi kuukautta, jotta trajektoreista saa muutakin esiin kuin vakioita. Täten trajektorianalyysin kyky havaita lyhyen tähtäimen kehityksiä ennen mielenkiintoista tapahtumaa on heikko. Seuraava 4.13. kuvio esittää 61 päivän trajektoreita kolmannen polynomiasteen mallilla, jossa ryhmien koot ovat 101, 130, 207 ja 416 ja tiketit ovat 115, 147, 235 ja 486.

Näistä oikeanpuoleisin trajektori on selvästi suurin (49% osuus liittymistä ja 57% osuus tiketeistä) ja siinä on suhteessa hieman enemmän tikettejä kuin muissa ryhmissä. Tässä ryhmässä, joka samalla paljastaa yleisimmän käyttäytymismallin ennen tikettiä, paluusuunnan kohinamarginaali ja myötäsuunnan vaimennus ovat kasvavia muiden muuttujien ollessa vakioiset. Toiseksi suurimmassa ryhmässä paluusuunnan kohinamarginaali ja myötäsuunnan vaimennus ovat laskevia muiden muuttujien ollessa vakioiset ennen tikettiä. Toiseksi vasemmalla olevan ryhmän muut arvot ovat nousevia paitsi myötäsuunnan kohinamarginaali on laskeva. Vasemmanpuoleisin, samalla pienin ryhmä edustaa harvinaisempaa kehityskulkua ennen tiketin ilmaantumista.



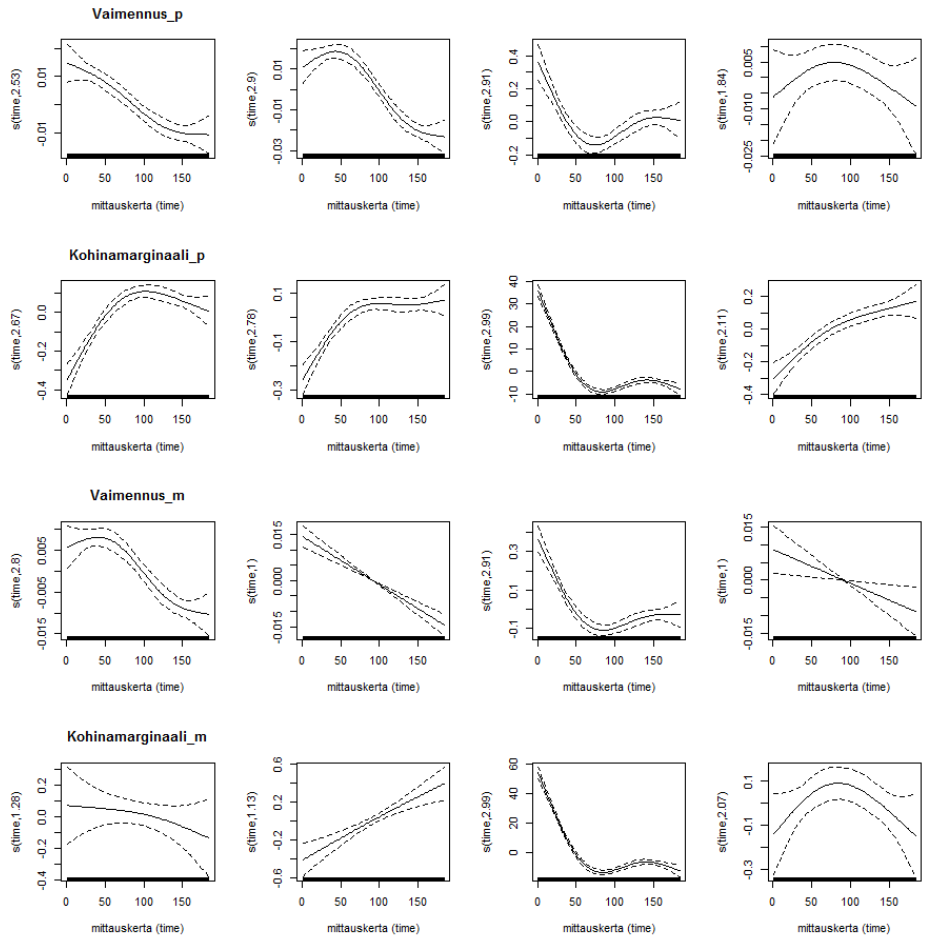

**Kuvio 4.13.** Katkaistut tiketilliset liittymät 61 päivän aikaikkunalla.

## 4.6 Liittymien lukumäärän vaikutus malliin

Edellä tehtyjen analyysien perusteella liittymien lukumäärä on ratkaiseva trajektorianalyysissä saatujen ryhmien lukumäärän suhteen määrittäen sitä kautta myös trajektoreiden mahdollisia muotoja ja yhdistelmiä. Tiketillisten liittymien määrä on 732, jolla onnistui saada monipuolisesti erilaisia ryhmämääriä. Sen sijaan aineiston A sisältämät 1811 liittymää tarjosivat korkeintaan kolme ryhmää, mikä vaikuttaa jatkoanalyysien perusteella liian vähältä ollakseen tulosten osalta niin hyödyllinen. Samalla trajektoreiden muodot olivat kovin pelkistettyjä.

### 4.6.1 Aineiston A trajektoreita eri liittymämäärillä

Seuraavaksi on haarukoitu, millä maksimi liittymämäärällä aineisto A mahdollisesti tarjoaisi monipuolisemmin ryhmiä tarkastelemalla aineistosta otettuja satunnaisotoksia eri liittymämäärillä. Koeajoissa ilmeni, että noin 1000 liittymällä on mahdollista saada useamman ajon avulla ainakin neljän ryhmän tulos viidennen polynomiasteen mallilla. Kuitenkin vasta noin 300 liittymän lukumäärä on sellainen, joka tarjoaa useamman ajon jälkeen myös viisi ryhmää. Viisi ryhmää jakaa aineiston vähät tiketit useammalle ryhmälle, joten se ei vaikuttanut enää niin mielenkiintoiselta. Noin 100 liittymää vaikutti olevan pienin liittymämäärä, jolla saa muutakin kuin vakioisia trajektoreita. Seuraavassa kuviossa on esitetty 900 liittymän neljän ryhmän ja viidennen polynomiasteen mallin monitrajektorit, joissa liittymien lukumäärät ovat 197, 251, 56 ja 396 ja tiketit vastaavasti 3, 10, 0, ja 18.



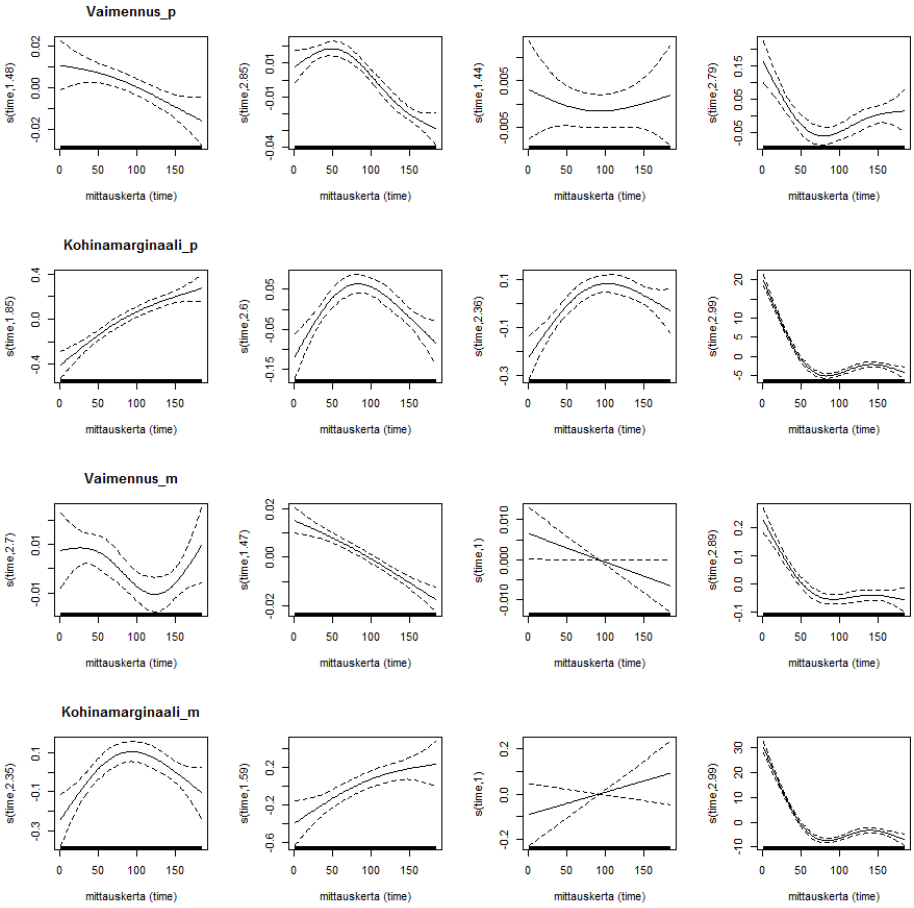

**Kuvio 4.14.** 900 liittymän monitrajektorit.

Koska aineistossa A on sekä tiketillisiä ja ilman tikettejä olevia liittymiä, tarvitaan riittävä määrä ryhmiä, joiden joukossa on mahdollisesti yksi tai useampi

ryhmä, joka kerryttää suhteellisesti selvästi enemmän tikettejä kuin muut. Kuviossa 4.14. näitä ovat samalla kooltaan suurimmat oikeanpuoleisin ja toinen vasemmalta sijaitsevat ryhmät. Kahdessa muussa ryhmässä tikettien määrä on varsin pieni.

### 4.6.2 Tiketittömät liittymät

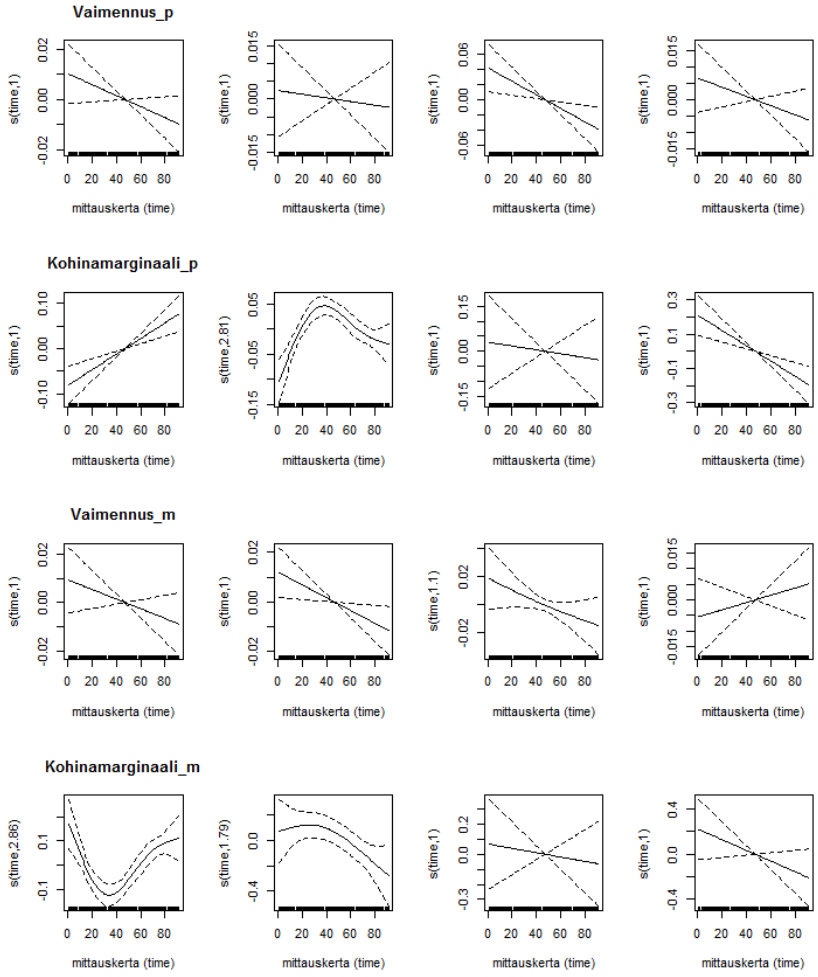
Aineistosta A otetaan vielä satunnaisotos, jossa on vain liittymiä, joilla ei ole yhtään tikettiä. 700 liittymällä saadaan neljän ryhmän tulos, josta esimerkki kuviossa 4.15. Liittymien määrät ryhmittäin ovat 280, 197, 127 ja 5. Kun näitä trajektoreita vertaa tiketillisten liittymien ryhmiin, niin selvää yhtäläisyyttä on havaittavissa pareittain. Kaikkien neljän trajektorin perusmuodot ovat samat ainoastaan ryhmäparilla, joka sisältää molempien mallien pienimmän ryhmän. Näin ollen tiketillisillä liittymillä näyttäisi olevan hieman erilaiset monitrajektorit verrattuna tiketittömiin liittymiin noin samankokoisilla otoksilla pääerojen kohdistuessa lopuilla kolmella ryhmällä 1 - 2 eri trajektorin muotoon.




**Kuvio 4.15** Tiketittömien liittymien monitrajektorit 700 liittymällä.

### 4.6.3 Katkaistut tiketittömät liittymät

Aineiston A tiketittömistä liittymistä tehdään 61 päivän ja 732 liittymän katkaistu aineisto, jossa viimeinen päivä on aineistossa oleva viimeinen mittaus. Saatuja monitrajektoreita verrataan vastaavalla tavalla katkaistujen tiketillisten liittymien monitrajektoreihin (luku 4.5.3). 61 päivän aikaikkunalla trajektorit osoittautuvat pääosin vakioisiksi, joten käytetään pitempää 91 päivän katkaistua aikasarjaa, jonka monitrajektorit on esitetty kuviossa 4.15. Ryhmien koot ovat vasemmalta oikealle 214,176, 173 ja 169.

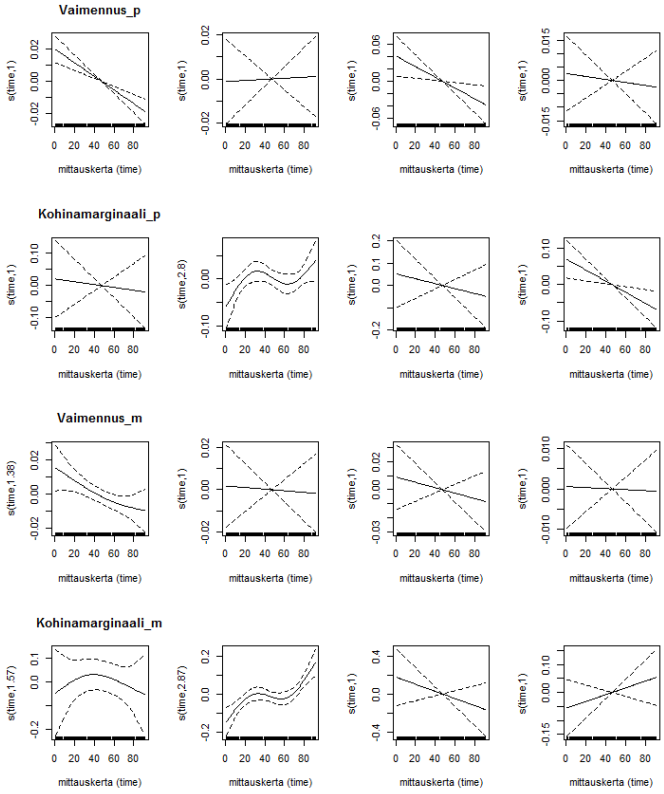



**Kuvio 4.16.** Katkaistut tiketittömät liittymät 91 päivän aikaikkunalla.

Kun vertaa yllä olevaa tulosta katkaistujen (kuvio 4.13) tiketillisten liittymien monitrajektoreiden kanssa, niin ryhmien välille ei löydy suoria vastineita. Tiketillisillä liittymillä paluusuunnan kohinamarginaali ja myötäsuunnan vaimennus ovat aina muuta kuin vakiot; tiketittömillä kaikki myötäsuunnan vaimennukset ovat vakiot ja yhdessä ryhmässä myös paluusuunnan kohinamarginaali on vakio.

### 4.6.4 Katkaistut tiketilliset ja tiketittömät liittymät

Aineiston A liittymistä tehdään 91 päivän ja 732 liittymän katkaistu aineisto, jossa viimeinen päivä tarkoittaa aineistossa olevaa viimeistä mittausta. Saatuja monitrajektoreita verrataan vastaavalla tavalla katkaistujen tiketittömien liittymien monitrajektoreihin (luku 4.6.3). Saadut monitrajektorit on esitetty kuviossa 4.17. Ryhmien koot ovat vasemmalta oikealle 217, 193, 171 ja 151 ja tiketit ovat 6, 9, 2 ja 3. Monitrajektoreiden muodossa on samankaltaisuutta tiketittömien trajektoreiden kanssa, mikä ei sinänsä yllätä, sillä vain osa liittymistä on vaihtunut näiden mallien kesken. Tämäkin pieni vaihdos on aiheuttanut sen, että vakioisissa muuttujissa tapahtuu jonkin verran muutoksia 95% luottamusvälillä. Aikaikkunan nosto 121 päivään ei tuonut lisää ymmärrystä tilanteeseen.



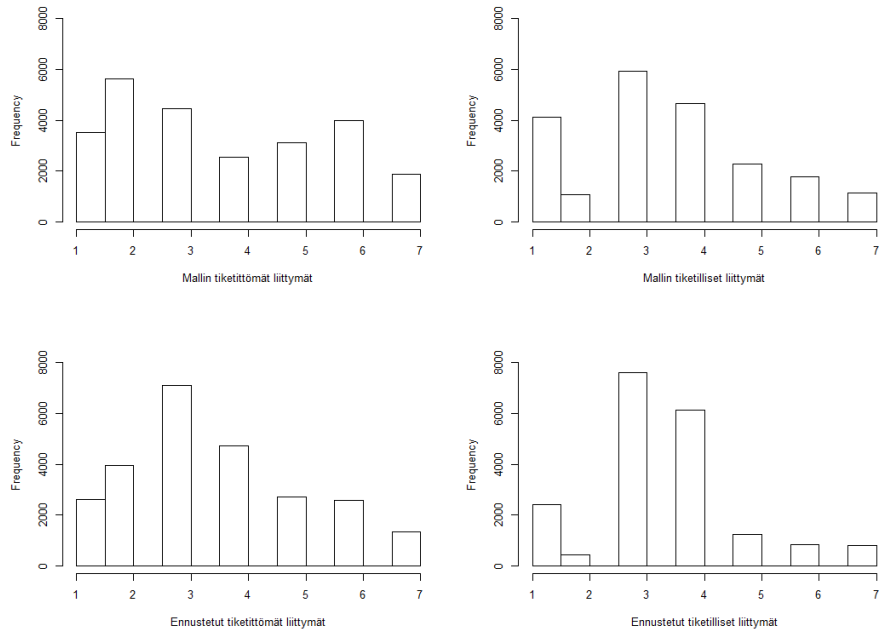

**Kuvio 4.17.** Katkaistut aineiston A liittymät 91 päivän aikaikkunalla.

Koska aineistossa A on suhteessa vähän tiketillisiä liittymiä, tehdään vielä aineiston A ja B avulla 61 päivän katkaistu 1708 liittymän testiaineisto, jossa on puolet ja puolet tiketillisiä ja tiketittömiä liittymiä. Edelleen tiketit jakautuvat kaikkien ryhmien välille tikettien määrän suhteessa liittymiin vaihdellen 7 - 51% välillä. Täten tiketillisten liittymien pieni suhteellinen osuus ei selitä sitä, että tehdyt ryhmitykset eivät erottele riittävän hyvin tiketillisiä ja tiketittömiä liittymiä eri ryhmiin. Lisäksi trajektorianalyysien ryhmien määrä vaihtelee neljästä kuuteen. Täten koeotoksessa on aineisto A:ta enemmän vaihtelevuutta, mitä ku-

vaamaan tarvitaan useampi ryhmä. Tämäkin kertoo siitä, että on hyvin käytetystä aineisto-osasta riippuvaista, kuinka paljon ryhmiä saadaan analyysin lopputulokseksi eikä liittymämäärä yksistään ole ratkaiseva tekijä ryhmälukumäärän määrittymisessä.

## 4.7 Trajektorimallin avulla ennustaminen

Tehdään ensin kolmannen polynomiasteen ja seitsemän (mahdollisimman suuri ajoissa tehty lukumäärä) ryhmän malli (monitrajektorit liitteessä E) 61 päivän katkaistuille 854 liittymille, joista puolet on tiketillisiä. Tämän jälkeen ennustetaan, miten mallin aineistoon kuulumattomien 427 tiketillisen ja tiketittömien liittymien mittaukset ryhmittyvät saatuun malliin. Ennustettavat tiketillisten liittymien mittaukset ryhmittyvät suhteessa selvästi eniten ryhmiin 3 ja 4, mutta samat ryhmät ovat suurimmat myös tiketittömien liittymien mittauksilla (kuvio 4.18.). Täten ennustaminen ei tuota riittävän tarkkoja tuloksia ollakseen hyödyllinen.




**Kuvio 4.18.** Liittymien ryhmittyminen: ylärivillä mallin liittymien ryhmät ja alarivillä ennustetut uusien liittymien ryhmät, ensin tiketittömät ja sitten tiketilliset liittymät.

# 5 Johtopäätökset

Reaaliaikaisten prosessien tuottaman pitkittäisaineiston käyttö toiminnan ohjaamiseen edellyttää, että analyyseissä käytetty data tulee kerättyä vaadittavalla laatutasolla. Tämä tarkoittaa käytännössä sitä, että datan keruuprosessin tuloksia pitää seurata sekä päivätasolla että pitemmällä aikajänteellä. Erilaisista syistä mittauksia jää paikoitellen saamatta ja jos tilannetta ei seurata pitkäjänteisesti, osa puuttuvasta tiedosta voi kroonistua. Lisäksi useammasta lähteestä kerättävän datan muoto, sisältö ja käytettävyys automatiikassa on ratkaisevassa roolissa analytiikan tehostajana. Tiedonhallinnan kokonaisarkkitehtuurin tehtävä on mahdollistaa hyvä analytiikka oli se rooli lopulta ihmisen, automatiikan tai molempien vastuulla.

Lopputyössä tutkittiin trajektorianalyysin mallinvalintaan liittyviä erilaisia seikkoja. Polynomiasteen valinnalla on mahdollista vaikuttaa saatavien ryhmien lukumäärään. Mitä suurempi polynomiaste on käytössä, sitä suurempi ryhmien lukumäärä saadaan tulokseksi aineistokohtaiseen rajaan saakka. Tämän jälkeen ryhmien lukumäärä jälleen pienenee. Ryhmien lukumäärä vaikuttaa trajektorianalyysin kestoon: mitä suurempi ryhmälukumäärä saadaan, sitä kauemmin analyysiajon suoritus kestää. Samalla ryhmälukumäärällä saavutetaan joko täysin sama tai hyvin samanlainen tulos riippumatta käytetystä polynomiasteesta.

Tyypillinen trajektorianalyysin haaste on päättää sopiva ryhmien lukumäärä. Lukuisten analyysiajojen perusteella toimialakohtainen ymmärrys vaikuttaisi parhaalta kriteeriltä valita oikea ryhmälukumäärä valittuun tarkoitukseen. Tämän jälkeen BIC-arvoa voi käyttää saman ryhmälukumäärän tuottamien eri mallien välillä auttamaan lopullisen mallin valinnassa. Ryhmälukumäärä itsessään ei ole kuitenkaan riittävä valintakriteeri. Monitrajektoreiden tuloksissa on ratkaisevaa se, kuinka sopiva yhtäaikaisesti tarkasteltavien mittausyksiköiden lukumäärä on. Aivan aluksi onkin syytä etsiä sopivan suuri aineistokokonaisuus, jolla trajektoreiden laadullinen sisältö yhdessä ryhmälukumäärän kanssa vastaa toimialaymmärrystä. Lopputyössä tutkitulla aineistolla päästiin parhaisiin tuloksiin 100 - 1000 liittymän aineistoilla. Mahdollisimman lyhyen mutta kuitenkin vielä hyödyllisen tarkastelujakson käyttö lyhentää suoritusaikoja ja minimoi vaadittavan tiedon tarpeen. Liian lyhyt mittausjakso tuottaa pääosin vain vakioarvoisia trajektoreita, mihin auttaa parhaiten aikasarjan pidentäminen, kunnes tulokset ovat mielenkiintoisempia.

Laajakaistaverkon aineistolle tehtyjen analyysien tulokset ovat todiste siitä, että aineistosta on löydettävissä ryhmiä, joilla on erilaisia kehityskulkuja. Osa kehityskuluista johtaa todennäköisemmin liittymän vikaantumiseen kuin toiset. Lopputyössä käytetyissä trajektorianalyyseissä on aina käytetty neljää muuttujaa, paluu- ja myötäsuunnan vaimennus ja kohinamarginaali, jotka on valittu toimialatietämyksen ja ennen lopputyötä tehtyjen esitutkimusten perus-

teella. Vaikka trajektorianalyysin käyttö vikaantumisen hallinnassa ei osoittautunut niin hyödylliseksi kuin lähtöodotus oli, analyysien avulla saatiin kuitenkin lisätietoa suurista käyttäytymisen linjoista eri tavoin kuin aiemmin; tiedon tiivistäminen trajektorianalyysin avulla toimii.

Tarkasteltaessa erikseen tiketittömiä ja tiketillisiä liittymiä, trajektoreissa on merkittäviä eroja erityisesti myötäsuunnan vaimennuksen ja paluusuunnan kohinamarginaalin kohdalla. Tiketillisillä liittymillä kehityskulut ovat näissä selvästi laskevia tai nousevia ja tiketittömillä pääosin vakioiset. Valitettavasti, kun tarkastelee aineiston osaa, joka sisältää todellisilla tai tarkoitushakuisesti muodostetuilla suhteilla sekä tiketittömiä että tiketillisiä liittymiä, vastaavia monitrajektoreita ei saada yhtä aikaa esiin, jotta ennakointia voisi hyödyntää tuloksellisemmin. Pääsyy tähän lienee se, että ryhmittymisen pääkriteerit eivät ole yhtenevät tiketöitymiseen johtavien syiden kanssa tai nämä tikettiin johtavat syyt ovat niin hienosyiset, että ne eivät erotu riittävällä tavalla ryhmittymisessä ainakaan lopputyössä valittujen neljän muuttujan avulla. Lisäksi vaikuttaa siltä, että tutkitulla aineistolla ei ole mahdollista saada informatiivisia lyhyenajan trajektoreita; käytännössä noin kolme kuukautta osoittautui lyhyimmäksi hyödylliseksi tarkastelujaksoksi. Täten ainakin tällä aineistolla tutkitulla menetelmällä on mahdollista tarkastella pääosin vain pitemmän ajan kehityskulkuja.

On kuitenkin hyvä muistaa, että esitellyt analyysit ovat vain näytteitä valituilla otoksilla ja muuttujilla. Todennäköistä on, että myös muunlaisia multitrajektoreita esiintyy, kun vain riittävästi etsii. Ja jos löytyy riittävän hyödyllinen malli, sitä kannattaa käyttää mittausten seurannassa ja ennakoinnissa, kunnes löytyy parempi. Nyt ei kuitenkaan löytynyt sellaista mallia, jota voisi suoraan suositella käyttöönotettavaksi. Koska tuotantoympäristössä ei voi nojata mittauspäivän tiketteihin, niin sopivalla aikaikkunalla katkaistu mittausaineisto päättyen aina kyseiseen mittauspäivään vaikuttaisi parhaalta tavalta ennakoida mahdollisia vikatilanteita. Tämä vaatii kuitenkin nyt puuttuvan hyödyllisen mallin monitrajektoreineen. Automatiikan käyttöön tarkoitetun mallin ryhmälukumäärä ei ole niin rajallinen kuin jos ihminen olisi mallin käyttäjänä. Täten ennakointia varten tehty malli lienee parasta koostaa mahdollisimman rikkaan aineiston avulla, jonka avulla on mahdollista saada paljon ryhmiä, joiden avulla aineiston sovitus olisi hyvällä tasolla. Mallin hyödyllisyys määrittyy lopulta ennakoinnin osuvuudella. Koska trajektorianalyysilla ei voi tarkastella yhtäaikaisesti koko suurta aineistoa, on oletettavaa, että myös tulokset vaihtelevat jossain määrin siirryttäessä aineiston osasta toiseen, vaikka mallin muuttujat ja toimialue ovat samat.

Aikasarjagraafien, jotka muuntavat läpinäkemättömän mustan laatikon läpinäkyväksi, välitön integrointi vianhallintaan on hyödyllistä, sillä niiden avulla on mahdollista nähdä tiiviissä muodossa tarkasteltavana olevan liittymän tila korkeintaan vuorokauden viiveellä. Tätä ennen liittymän toiminnan ymmärrys on ollut kovin pirstaleista ja on nojannut pitkälti tiedon käyttäjän kykyyn muodostaa kokonaiskuva erilaisista tiedonpalasista. Aikasarjagraafeja voidaan käyttää myös vianhallinnan kehittämis- ja testaustyökaluna, sillä niiden avulla voidaan todentaa tehtyjen muutosten vaikutuksia uudella tavalla.

Soveltavien analyysien aikana selvisi, että käytössä olevassa aineistossa ei ole todennäköisesti lainkaan kaapeleiden kastumiseen liittyviä mittauksia, jo-

ten kyseistä ilmiötä ei voinut tarkastella mitenkään tällä erää. Pitemmän aikavälin vikaantumisilmiönä se olisi kuitenkin mielenkiintoinen jatkotutkimuskohde, mikäli sopiva aineisto vain on käytettävissä. Tutkimuksen aikana kävi myös selville, että liittymien mittausten lukumäärät vaihtelevat suuresti yhden mittauksen ja koko tarkastelujakson päivien lukumäärän välillä. Oletettavaa on, että tämä tuo ylimääräistä epätarkkuutta tuloksiin, ja olisi mielenkiintoista tutkia, mikä olisi sopiva vaatimus analyysiin otettavien havaintojen olemassaolon tasolle. Nyt kaikki liittymät on sisällytetty mukaan analyyseihin huolimatta siitä, kuinka paljon mittauksia puuttuu, koska trajektorianalyysi ei vaadi tasapainoista dataa.

Trajektorianalyysi vaikuttaa lupaavalta menetelmältä, jolla on uudenlaista annettavaa reaaliaikaisen prosessin vianhallinnan automatisoinnissa. Menetelmän vahvuus on eritoten siinä, että sen avulla on mahdollista tarkastella aineistoa suhteessa aikaan aivan uudella tavalla useamman dimension kautta. Haasteellisinta lienee kuitenkin löytää hyödyllinen malli, joka vastaa käyttötarpeita riittävän hyvällä tasolla. Tässäkin tapauksessa tuotantoon sopivan mallin etsiminen jää jatkotutkimuksen varaan.

# Lähteet

Berg, N., Kiviruusu, O., Karvonen, S., Kestilä, L., Lintonen, T., Rahkonen, O. & Huurre, T. (2013), ”A 26-Year Follow-Up Study of Heavy Drinking Trajectories from Adolescence to Mid-Adulthood and Adult Disadvantage”, Alcohol and Alcoholism Advance Access.

Box, G. E. P., Cox, D. R. (1964), ”An Analysis of Transformations”, Journal of the Royal Statistical society, vol 26, no 2, saatavana: http://www.econ.uiuc.edu/~econ508/Papers/boxcox64.pdf

Dodge, H. H., Shen, C. & Ganguli, M. (2008), ”Application of the Pattern-Mixture Latent Trajectory Model in an Epidemiological Study with Non-Ignorable Missingness”, Journal of Data Science, vol 6, no 2.

D’Unger, A. V., Land, K. C., McCall, P. & Nagin, D. S. (1998), ”How Many Latent Classes of Delinquent/Criminal Careers? Results from Mixed Poisson Regression Analyses of the London, Philadelphia, and Racine Cohorts Studies.”, American Journal of Sociology, saatavana: http://www.jstor.org/discover/10.1086/231402?sid=21105782917471&uid=2&uid=4&uid=3737976

Eldén, L., Wittmeyer-Koch, L. & Nielsen, H. B. (2004), ”Introduction to Numerical Computation - analysis and Matlab illustrations”, Studentlitteratur AB.

Enders, C. K. (2010), “Applied Missing Data Analysis”, Guildford Press.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), ” Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal statistical Society, vol 39, no 1, saatavana: http://www.eng.auburn.edu/~troppel/courses/7970_2015A_AdvMobRob_sp15/literature/paper_Wrefs/dempster EM 1977.pdf

Faraway, J. J. (2006), “Extending the Linear Model with R – Generalized Linear, Mixed Effects and Nonparametric Regression Models”, Chapman & Hall/CRC.

Gildengers, A. G., Houck, P. R., Mulsant, B. H., Dew, M. A., Aizenstein, H. J, Jones, B. L., Greenhouse, J., Pollock, B. G. & Reynolds, C. F. (2005), ”Trajectories of Treatment Response in Late-Life Depression – Psychosocial and Clinical Correlates”, Journal of Clinical Psychopharmacology, vol 25.

Grün, B. & Leisch, F. (2008), “FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters”, The Comprehensive R Archive Network (CRAN), saatavana: ftp://videolan.c3sl.ufpr.br/CRAN/web/packages/flexmix/vignettes/mixture-regressions.pdf

Grün, B., Leisch, F. , Sarkar, D., Mortier, F. & Picard, N. (2015), "Package 'flexmix' – Flexible Mixture Modeling", The Comprehensive R Archive Network (CRAN), saatavana: http://cran.r-project.org/web/packages/flexmix/flexmix.pdf

Hasan, S. (2012), "Group based trajectories of network formation and dynamics", Elsevier B.V.

Higgins, G. E., Kirchner, E. E, Ricketts, M. L. & Marcum, C. D. (2013), "Impulsivity and Offending from Childhood to Young Adulthood in the United States: A Developmental Trajectory Analysis", International Journal of Criminal Justice Sciences, vol 8.

Hill, K. G., White, H. R., Chung, I., Hawkings, J. D. & Catalano, R. F. (2000), "Early Adult Outcomes of Adolescent Binge Dringking: Person- and Variable-Centered Analyses of Binge Drinking Trajectories", Alcoholism: Clinical and Experimental Research, vol 24, no 6.

Hyppänen, J. (2007), "Ennakoiva vikojen havaitseminen liityntäverkosta", Lappeenrannan teknillinen yliopisto, saatavana: http://www.doria.fi/handle/10024/30005

Jones, B. L. & Nagin, D.S (2007), "Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them", Sociological Methods & Research, vol 35, no 4.

Kass, R. E. , Raftery, A. E. (1995), "Bayes Factors", Journal of the American Statistical Association, vol. 90, no. 430, saatavana: http://xyala.cap.ed.ac.uk/teaching/tutorials/phylogenetics/Bayesian_Workshop/PDFs/Kass and Raftery 1995.pdf

Leisch, F. (2003), "FlexMix: A general framework for finite mixture models and latent class regression in R", saatavana: http://epub.wu.ac.at/712/1/document.pdf

McLachlan, G. &Peel, D. (2000), "Finite Mixture Models", John Wiley & Sons.

Matthews, J. W. (2015), " Group-based modeling of ecological trajectories in restored wetlands", Ecological Applications, vol 25, no 2.

Nagin, D. S. (1999), "Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach", Pyschological Methods, vol 4, no 2, 139-157.

_____ (2005), "Group-Based Modeling of Development", Harvard University Press.

Muthén, B. (1989), "Latent Variable modeling in heterogenous populations", Psychometrika, vol 54, 557-585.

_____(2004), "Latent Variable Analysis – Growth Mixture Modeling and Related Techniques for Longitudinal Data - The SAGE Handbook of Quantitative Methodology for the Social Sciences edited by David Kaplan", Sage Publications, luku 19, 345-317.

Nagin, D.S. & Odgers, C. L. (2010), “Group-Based Trajectory Modeling in Clinical Research”, Annual Review of Clinical Psychology, 6, 109-138.

Nagin, D. S. & Tremblay, R. E. (2001), “Analyzing Developmental Trajectories of Distinct but Related Behaviours: A Group-Based Method”, Psycological Methods, vol 6, no 1, 18-34.

Nummi, T., Hakanen, T., Lipiäinen, L., Harjunmaa, U., Salo, M. K., Saha, M.,Vuorela N. (2013), ”A trajectory analysis of body mass index for Finnish children”, Journal of Applied Statistics.

Roeder, K., Lynch, K. G.& Nagin, D. S. (1998), ” Modeling uncertainty in latent class membership: A case study in criminology”, Carnegie Mellon University, saatavilla: http://www.stat.cmu.edu/tr/tr683/tr683.html

Sampson, R. J., Laub, J. H. & Eggleston, E. P. (2004), ”On the Robustness and Validity of Groups”, Journal of Quantitative Criminology, vol 20, no 1.

Schwarz, G. (1978), ”Estimating the dimension of a model”, The Annals of Statistics, vol 6, no 2, saatavana: http://projecteuclid.org/download/pdf_1/euclid.aos/1176344136

Warner, L. A., White, H. R. & Johnson, V. (2007), ”Alcohol Initiation Experiences and Family History of Alcoholism as Predictors of Problem-Dringking Trajectories”, Journal of Studies on Alcohol and Drugs.

Virtanen, P., Lipiäinen, L., Hammarström, A., Janlert, U., Saloniemi, A & Nummi, T. (2011), ”Tracks of labour market attachment in early middle age; A trajectory analysis over 12 years”, Advances of Life Course Research, vol 16.

Wood, S. (2015), ”Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation”, The Comprehensive R Archive Network (CRAN), saatavana: http://cran.r-project.org/web/packages/mgcv/mgcv.pdf

# Liitteet

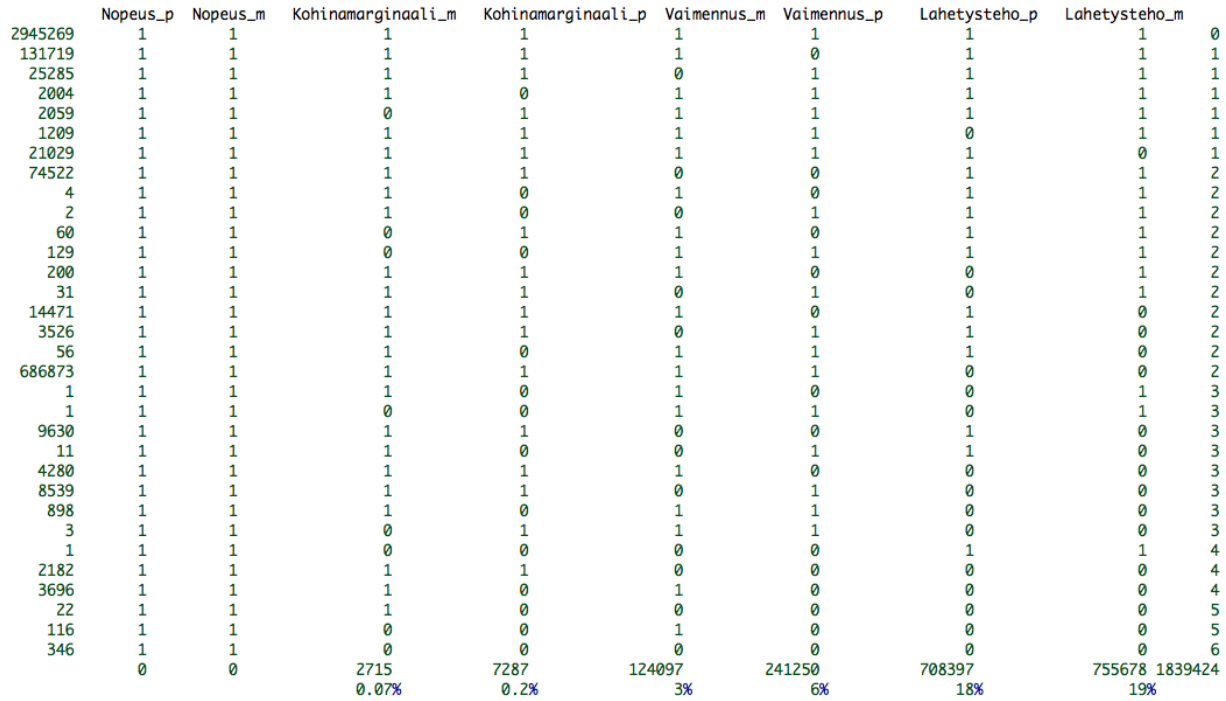
## Liite A: Mittausaineiston muuttujat

| Muuttuja | Kuvaus | Lisätietoa |
|---|---|---|
| ID | Liittymän yksilöivä tunniste | vakio |
| Mittauspv | Mittausrivin luontipäivä | 1.8.2014-31.1.2015 |
| Profiili | Asiakkaalle provisioitu liittymätyyppi (kättely+nopeus) | luokittelutyyppi |
| Alue | Laajempi maantieteellinen alue, joka sisältää useamman DSLAM:n hallinnoimat liittymät (esim. kaupunki) | Luokittelutyyppi |
| Keskus | Yhden DSLAM-keskuksen hallinnoimat liittymät | Luokittelutyyppi |
| Nopeus_p | Paluusuunnan (up) realisoitunut suurin mahdollinen nopeus | |
| Nopeus_m | Myötäsuunnan (down) realisoitunut suurin mahdollinen nopeus | |
| Vaimennus_p* | Paluusuunnan realisoitunut vaimennus | dB |
| Vaimennus_m* | Myötäsuunnan realisoitunut vaimennus | dB |
| Kohinamarginaali_p** | Paluusuunnan realisoitunut kohinamarginaali | dB |
| Kohinamarginaali_m** | Myötäsuunnan realisoitunut kohinamarginaali | dB |
| Lahetysteho_p*** | Paluusuunnan realisoitunut lähetysteho | |
| Lahetysteho_m*** | Myötäsuunnan realisoitunut lähetysteho | |
| Tiketit | Mittauksiin liittyvien käyttäjän avaamien vikaselvitysten lukumäärä | |

* Vaimennus kuvaa sitä, kuinka paljon signaali vaimenee DSLAM:n ja DSL-modeemin välillä. Mitä alempi vaimennuksen desibeli-arvo on, sitä parempi tulos.
** Kohinamarginaali kuvaa DSL-signaalin ja linjalla olevan kohinan suhteellista voimakkuutta (Signal to noise ratio, SNR). Mitä suurempi kohinamarginaalin desibeli-arvo on, sitä parempi tilanne. Kaapelin pituus vaikuttaa ratkaisevasti kohteena olevan laitteen vaimennuksen ja kohinamarginaalin tasoon.
*** Lähetysteho kertoo, kuinka paljon laite kuluttaa tehoa sen kautta kulkevan tietoliikenteen palvelemiseen myötä- ja paluusuuntiin.

## Liite B: Mittausten puuttuneisuus

Alkuperäisen aineiston muuttujien puuttuneisuuden rakenne:



| | Nopeus_p | Nopeus_m | Kohinamarginaali_m | Kohinamarginaali_p | Vaimennus_m | Vaimennus_p | Lahetysteho_p | Lahetysteho_m | |
|---|---|---|---|---|---|---|---|---|---|
| 2945269 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 131719 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 25285 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 2004 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 2059 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1209 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 21029 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 74522 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 2 |
| 4 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 2 |
| 2 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 2 |
| 60 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 2 |
| 129 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 2 |
| 200 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 2 |
| 31 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 2 |
| 14471 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 2 |
| 3526 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 2 |
| 56 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 2 |
| 686873 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 2 |
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 3 |
| 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 3 |
| 9630 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| 11 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 3 |
| 4280 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 8539 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 3 |
| 898 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 3 |
| 3 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 4 |
| 2182 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 4 |
| 3696 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4 |
| 22 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 5 |
| 116 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 5 |
| 346 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| | 0 | 0 | 2715 | 7287 | 124097 | 241250 | 708397 | 755678 | 1839424 |
| | | | 0.07% | 0.2% | 3% | 6% | 18% | 19% | |

Alkuperäisen aineiston (AA) ja täydellisten mittausten (TM) tunnusluvut:

| | | Nopeus_p | Nopeus_m | Kohinamar-mar-ginaali_p | Kohinamar-mar-ginaali_m | Vaimen-nus_p | Vaimen-nus_m | Lahetys-teho_p | Lahetyst eho_m |
|---|---|---|---|---|---|---|---|---|---|
| AA | Mediaani | 1019 | 2464 | 9.1 | 18.0 | 8.0 | 14.5 | 108.0 | 163.0 |
| | Keskiarvo | 2245 | 6407 | 11.8 | 20.4 | 10.3 | 17.3 | 98.3 | 135.8 |
| TM | Mediaani | 1019 | 2464 | 8.8 | 20.6 | 7.1 | 17.0 | 112.0 | 171.0 |
| | keskiarvo | 1450 | 6653 | 11.2 | 21.8 | 9.5 | 19.4 | 100.3 | 141.8 |

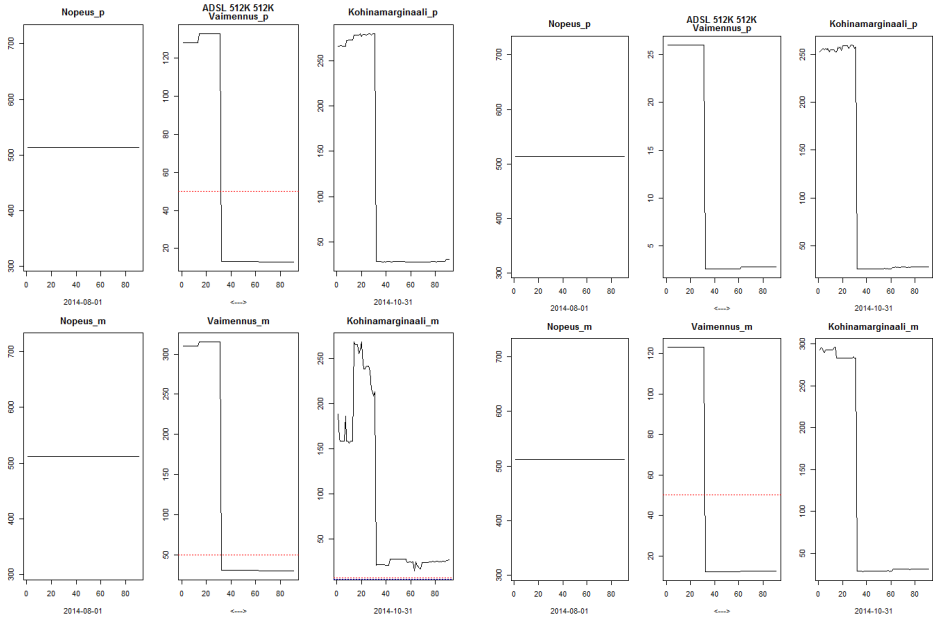
## Liite C: Raaka-datan laadullisia ominaisuuksia

- Kategorisen muuttujan arvot tulee olla ennalta valitut, eikä käyttäjän omia käsin kirjoitettuja vaihtoehtoja tule sallia. Samaa asiaa tarkoittavia valintavaihtoehtoja on vain yksi kerrallaan eikä eri kieliversioita sallita. Mahdollinen kielituki tulee hoitaa muuta kautta.
- Muuttujan kentän arvot tulee olla aina samassa yksikössä eikä saa olla mitään toisen muuttujan arvon kautta tapahtuvaa tulkintaa. Automatiikka ei pysty suoraan käsittelemään tällaista tietoa vaan tarvitaan ylimääräistä logiikkaa. Lisäksi ilman automatiikkaa tällaisen tiedon käsittely on kuormittavaa, virheherkkää ja tehotonta.
- Samaa muuttujaa tarkoittavien kenttien nimet olisi hyvä olla yhtenevät kautta organisaation, kun kyseessä on sama asia. Tämä minimoi samalla työntekijöiden kuormittumista, kun tutkitaan useammasta lähteestä tullutta tietoa ilman automatiikkaa.
- Perustietojen oikeellisuus ja täydellisyys on oltava korkealla tasolla, kun tietoa käytetään automatiikkaan. Analyyseissä ei voi hyödyntää sellaista tietoa, jonka tulkinta on epämääräistä, vaarantamatta koko analyysin laatua. Turhaa tiedon suodatusta tulee välttää näistä syistä, kun asiat on mahdollista laittaa kuntoon huolellisella tiedonhallinnalla.

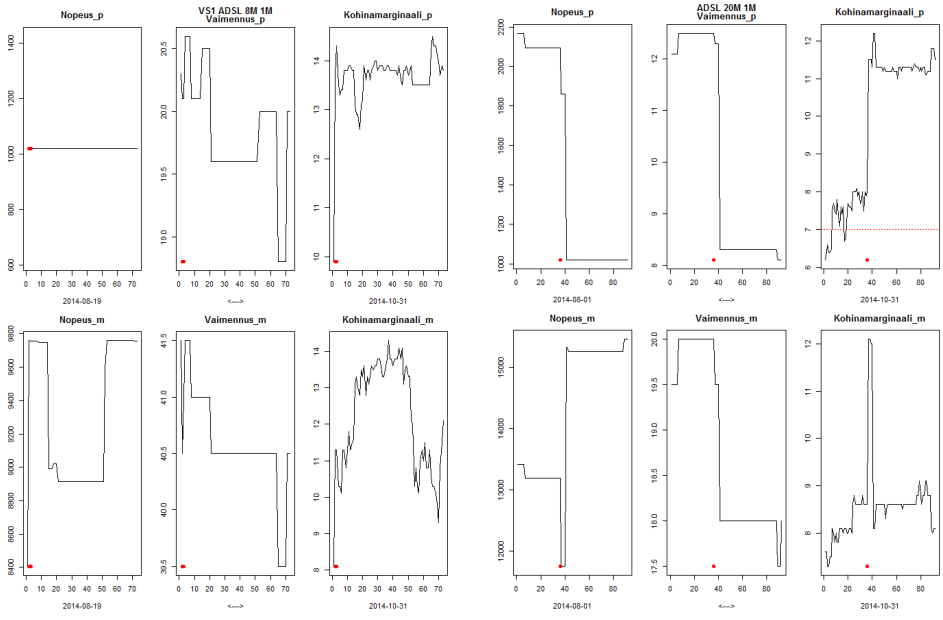
# Liite D: Kolmen kuukauden ryhmien aikasarjoja

Oheiset kuviot sisältävät kuviossa 4.8 esitettyihin kolmeen ryhmään kuuluvien liittymien aikasarjagraafeja.

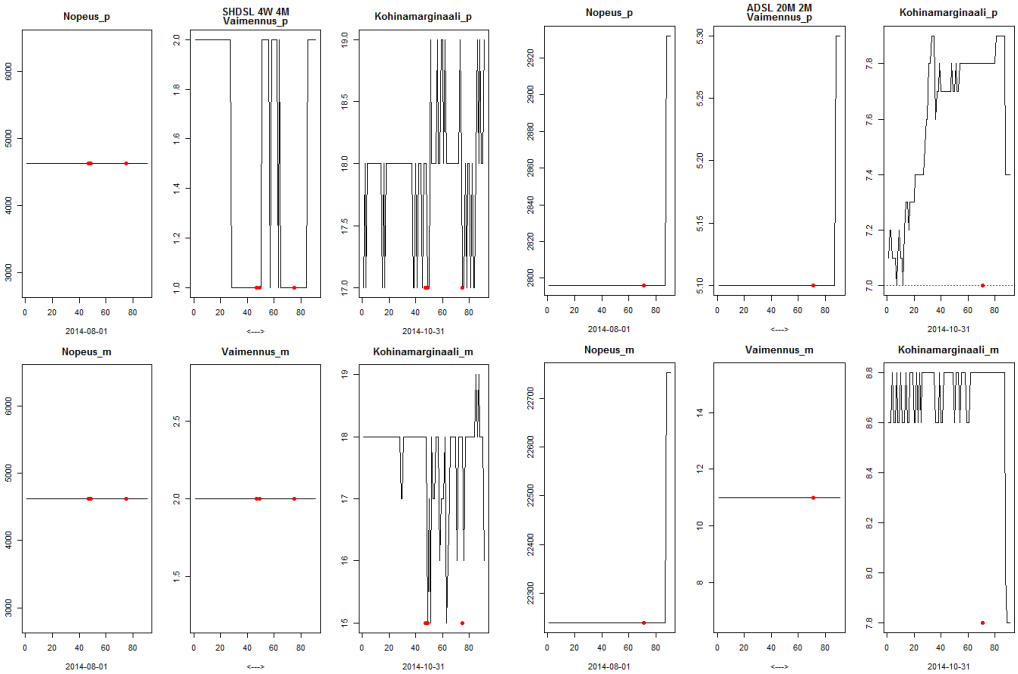
Vasen ryhmä:




Keskimmäinen ryhmä:



Oikeanpuoleinen ryhmä:



# Liite E: Ennustuksessa käytetyn mallin monitrajektorit